Estimating Train Passenger Load from Automated Data Systems: Application to London Underground

By

Elizabeth Cheriyamadam Paul

Bachelor of Arts, Mathematics

City University of New York Hunter College, 2008

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

Massachusetts Institute of Technology

September 2010

Signature of Author...

Department of Civil and Environmental Engineering

June 1, 2010

Certified by...

Nigel H. M. Wilson

Professor of Civil and Environmental Engineering

Thesis Supervisor

Certified by...

John P. Attanucci

Research Associated of Civil and Environmental Engineering

Thesis Supervisor

Accepted by...

Danielle Veneziano

Chairman, Departmental Committee for Graduate Students

Estimating Train Passenger Load from Automated Data Systems: Application to London Underground

By
Elizabeth Cheriyamadam Paul

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Transportation
at the
Massachusetts Institute of Technology

September 2010

ABSTRACT

The purpose of this thesis is to assess the feasibility of identifying which trains individual passengers take to get from their origin to destination while travelling in a high frequency urban rail transportation system. If this proves possible, the resulting information will inform capital and operations planning decisions as well as improve the ability to measure the aspects of passenger experience related to travel time and crowding. This thesis will first explore this idea by presenting the design, implementation, and application of a model that attempts to identify the selected train level itineraries through a temporal and spatial matching process. As a result of this process, the model is designed to estimate passenger loads, walk times, and the number of left behind passengers. The thesis will then assess the accuracy of these results by comparing them with figures produced by existing models. The model will be developed and applied in the context of the London Underground, but should also be applicable to other urban public transportation systems. Assessment of the results of this model and consideration of the challenges in the creating the model does not conclusively indicate that identifying the exact train a passenger selects to get from his origin to destination is possible. However, the results do indicate that the model has significant potential, and can be improved in future research. These initial results can serve as indicators on how to improve the model.

Thesis Supervisor: Nigel H.M. Wilson
Title: Professor of Civil and Environmental Engineering

Thesis Supervisor: John P. Attanucci
Title: Research Associate of Civil and Environmental Engineering

Acknowledgements

I would like to thank my research advisors Professor Nigel Wilson and John Attanucci. I have learned a great deal over these past two years from them and cannot thank them enough for this opportunity. I will never forget their dedication to their students and the quality of research.

Huge thanks also go to the first year students who have aided me in my research: Sam Hickey and Matt Shireman. It has been such a pleasure to work with both of you. This thesis wouldn't be possible without you.

Thank you to the folks at Fares and Ticketing and S&SD for taking the time to supply me with information, suggestions, and useful feedback. Special thanks to Nigel Kelt for the supervision, frank comments, and hours of programming lessons. I hope we can work together again in the future.

Next, I would love to thank my fellow MSTs for laughing and brainstorming with me, and giving me company as the delirium set in. Valerie, my amazing and thoughtful friend, thank you for the source of so many fun memories during my time here in Boston. Julian, my favorite lunch buddy, thank you for always listening to me and helping me see the big picture. Thank you to Winnie for making me drink warm water (not), making me put less sugar in my coffee, and the random dancing. Thanks to Jared for the cookies, being my replacement Val, and being our chauffer (dad). I would also like to thank my roommate Sevara for being a great friend and not getting mad when the dishes sat in the sink for more than a week. Of course, thank you Candy, Andre and David for being awesome role models. And thanks to everyone else in the 1-235 that I haven't mentioned. It has been an incredible two years.

I would also like to thank my family: Mom, Dad, Deepti, Noel, Anna, and Jacob. Also my mentors in New York who got me interested in transportation and continue to support my endeavors: Charlotte Glasser, Hilda Montanez, Albert Chen, John Kennes, Charmaine Jordan, Roberta Glogover and Stacey Berman. I certainly wouldn't be here without you. Special thanks to Noel for the edits!

Thank you to my lovely ladies: Bridget, Ava and Rossana. The phone calls and gchats are much appreciated. Finally, I would like to thank Tom Henderson for the edits and encouraging me every step of the way through this thesis. Olives.

Table of Contents

List of Figures

1 Introduction

The purpose of this thesis is to assess the feasibility of identifying which trains individual passengers take to get from their origin to destination while travelling in a high frequency urban rail transportation system. If this proves possible, the resulting information will inform capital and operations planning decisions as well as improve the ability to measure the aspects of passenger experience related to travel time and crowding. This thesis will first explore this idea by presenting the design, implementation, and application of a model that attempts to identify the selected train level itineraries through a temporal and spatial matching process. As a result of this process, the model is designed to estimate passenger loads, walk times, and the number of left behind passengers. The thesis will then assess the accuracy of these results by comparing them with figures produced by existing models. The model will be developed and applied in the context of the London Underground (LU), but should also be applicable to other urban public transportation systems.

1.1 Motivation

Knowledge of urban rail systems usage and performance varies according to the technology available to these systems. Three measures of performance are of interest in this thesis: passenger load, walk time and wait time. These three measures can be estimated at different levels of aggregation. For example, average passenger load can be estimated at the line level in systems that have exit and entry fare control and minimal path choice within the network. In these systems, service is often represented by average line frequencies. In more complex networks, path choice must be modeled before average line loads can be estimated. Unfortunately, this high level of aggregation cannot provide passenger loads on individual trains. This knowledge would be useful in estimated passenger walk and wait time. In systems that have entry and exit fare control, and train movement data, there is potential for a greater degree of detail in measurement of performance. Passengers could be assigned to individual trains based on their entry and exit times and corresponding train movement records. Given this assignment, walk and wait times within stations could be computed for each passenger. Given individual wait times, it can be inferred whether a passenger is left behind by a departing train, either by choice or because of severe crowding. This potential for measuring individual train loads and passengers left behind using archived automated data is the motivation for this research.

Developing a working model of this type will facilitate the disaggregate analysis of passenger behavior in complex urban rail networks. It will allow for the examination and measurement of various aspects of individual passengers' experience between the time they enter and leave the transit network. The model aims to identify the specific train(s) a passenger boards, potentially producing detailed knowledge about the existence and effects of crowding and the allocation of passenger travel time between in train and out of train components. First, loads on each train on the transit system can be estimated. This can be used to assess the levels of crowding on each line in a transit system over time. Second, distributions of walk times can be created for each station in the transit system. Third, lines, stations and time periods in which passengers are left behind can be identified. A passenger is left behind when there is no available capacity on the train that he was in a position to board. Counting the number of left behind passengers by station or by train can help identify problem areas in the service currently provided, and also help identify the relationship between passenger loads on trains and the likelihood of passengers being left behind at a station.

1.2 Objective

The primary objective of this thesis is to test the feasibility of accurately assigning passenger journeys to individual trains given the times and locations of passenger entries and exits to a rail transit network and archived train movement records.

1.3 Research Approach

The above objective is the result of a logical sequence of increasing disaggregation of performance measures based on data available in modern transit systems.

As mentioned in section 1.1, a model of a transit system at a more aggregate level might represent services by average frequencies. This type of model can estimate average loads on trains, but cannot account for more subtle changes in service and demand, and cannot estimate passenger loads on individual trains (Nuzzolo, 2009). An improvement (in terms of detail) on this type of model is to represent service using train schedules. Load can be estimated on each individual train, and the effects of changes in schedule can be tracked. Schedule-based modeling of transit networks allows for more detailed analysis of effects of service on passengers. Both frequency and schedule-based models traditionally use demand data in the form of passenger surveys.

With the advent of smartcard data, the next logical step is to assign passengers to the scheduled trains. Since smartcard data reveals actual passengers' entry and exit time, these passengers can be assigned to scheduled trains in order to estimate the load on each train. Frumin (2010) proposes a method to study the relationship between the times of passenger arrival and departure, and published timetables by integrating disaggregate passenger journey data from automatic fare collection (AFC) systems with published timetables using schedule-based assignment. A drawback with this process is that trains do not always adhere to schedule. This means that even though a passenger may be assigned to a certain scheduled train, in reality, the passenger may not have boarded that train because it ran either later or earlier than scheduled. In this sense, the schedule-based modeling of transit networks falls short.

An improvement is to assign each passenger to an actual train run, as recorded by the transit network's train control or track signaling system. Using a model based on actual train runs is beneficial because it eliminates any infeasible assignments that may otherwise occur. Using smartcard data in conjunction with actual train runs allows for greater accuracy when estimating loads on individual trains. Furthermore, this type of model allows for the inference of actual passenger walk times, and the number of passengers left behind by trains, something that previous models did not allow.

The model developed in this research attempts to achieve the objective by matching smartcard data with train movement data from track signalling systems and by enumerating possible paths between origins and destinations on a transit network. The model developed in this thesis is designed to be adopted as a tool by any transit agency wishing to study the assignment of passengers to trains using smartcard data and train run data.

In this thesis, the model is developed and applied to the London Underground network, which is a subsidiary of Transport for London (TfL). Transport for London's smartcard system, "Oyster Card" is currently used as the fare medium for over 70% of all trips made within TfL and linked entry and exit Oyster Card transactions are used as a critical input to the model. Data on actual train runs is derived from LU's Network Management Information System (NetMIS), which tracks train movements on LU lines of service. The passenger data and train movement data are matched through a Route Choice Model that enumerates a set of possible paths for each relevant origin-destination pair.

Constraints affecting the input data create a set of challenges to the successful identification of the train itinerary each passenger ultimately takes. For example, Oyster is not used by all passengers travelling on the LU network. Also, NetMIS data is complete on only some LU lines. Limited passenger and train service data from London Underground mean that the entire LU network cannot be represented. Because of these limitations, this model is intended as a proof of concept and will focus on the areas in the LU network where there is complete data. The methodology outlines algorithms and procedures for producing complete output data once the input data is complete, as well as procedures for producing complete output data for current incomplete input data.

There is inherent uncertainty in identifying the correct train for each passenger even if there is complete data. The first and more general reason for this is that there is no way of determining which route between an origin and destination a passenger actually took. In many cases, multiple routes may be feasible for a passenger. Furthermore, within a single route, there may be multiple feasible trains for a passenger. Again, there is no definite way to determine which train the passenger took, and hence no way to validate that the assigned train was indeed the train taken by the passenger. Validation is a challenge because prior to this research, there have been no similar attempts to identifying the actual trains that passengers took.

In order to match passenger data with train service data, there is one basic requirement: that the Oyster system clocks are synchronized with the train movement clocks. Unfortunately, the datasets from London Underground indicate that this assumption is violated at certain stations.

The methodology presented in this thesis aims to achieve the objective while overcoming these challenges through a series of assumptions. London Underground's smartcard data from the AM Peak on a single day with few known disruptions in service is matched with actual train runs from NetMIS data on the same day. The Route Choice Model provides a set of possible routes between any origin and destination.

Finally, in order to assess how well the model performs, aggregate measures produced by this model are compared with corresponding measures from models in current use by LU.

1.4 Introduction to London Underground

The model developed in this thesis is designed for the London Underground system. The London Underground network is one of the largest rail transit systems in the world with a total length of 402 line-km, approximately half of which is below ground. It is comprised of 270 stations on 11 lines, with many of the lines having multiple branches.

Passengers are the focus of the London Underground system, indeed they are the reason that service is provided. This is reflected in the fact that the London Underground's performance metrics are centered around figures intended to describe the passenger experience on the Underground (Uniman, 2009).

The following subsections describe in detail a selection of current performance metrics and train simulation tools used by the London Underground that are relevant to this research.

1.4.1 Train Service Model (TSM)

TSM is a train simulation tool that uses an annual passenger travel survey for its passenger data and assigns passengers to scheduled train runs. TSM measures passenger loads on each scheduled train, and calculate the number of passengers left behind by each train. TSM's load measurements are relevant to this research because a different approach is taken: estimated passenger flow from a typical day is assigned to scheduled trains from a typical day to estimate load. TSM's method of calculating number of passengers left behind is also relevant to this research. TSM assumes that passengers are left behind once the passenger load on a particular train reaches the density of 5 people per square meter (Weston and Maunder, 1994). This relationship between percent of passengers left behind and train load can be represented by a step function that is 0% at all train loads up to the maximum load, beyond which all additional passengers are left behind. This thesis also attempts to measure the percent of passengers left behind. It hypothesizes that in reality, the relationship between percent of passengers left behind and train load is likely to be a more continuous function because passengers may have varying degrees of tolerance for crowding depending on personal preference, length of trip, and other factors

1.4.2 Journey Time Metric (JTM)

JTM is the primary performance metric for the London Underground. The Journey Time Metric compares the (estimated) journey times experienced by passengers with the scheduled travel time for the same journeys. The difference is the Excess Journey Time, which is used as an indicator for how well the system performed from the passenger's perspective. Journey times are divided into five components, each reflecting a different aspect of the journey: Ticket Purchase Time, Access, Egress and Interchange Time, Platform Wait Time, On Train Time, and Closures. For each component there is an actual and a scheduled value. One component that is relevant to this research is the Access, Egress and Interchange (AEI) sub-model, which compares the scheduled access, egress and interchange times (the amount of time it takes to walk through a station unimpeded by congestion) against measured and estimated walk times throughout the day. These measurements and estimations of walk times are relevant to this research because walk time distributions are potential outputs of the proposed model.

1.4.3 Route Choice Model (RCM)

A Route Choice Model (RCM) produced by the London Underground is the data source that enumerates all possible paths between every relevant Origin and Destination (OD) pair on the London Underground. It is a critical component of the model developed in this thesis and is used in several ways. This RCM can be divided into three parts. First is the Generation of Alternative Paths for each OD pair. Second is a Generalized Cost for each alternative for each OD pair which produces a ranking of alternatives in terms of attractiveness to passengers. Third is the choice probability for each alternative path in the set of paths for each OD pair. These choice probabilities are currently calculated by plugging in each alternative's utility functions (generalized cost) into a logit function (Weston, 2009).

The model developed in this thesis uses the first element, generation of alternative paths, to determine all the possible paths a passenger could have taken. Once all passengers are assigned to routes in this research, it is possible to compute choice probabilities for each alternative route that serves an OD pair. These probabilities can be compared to RCM's choice probabilities for each route and OD pair.

1.5 Thesis Organization

Chapter 2 reviews previous related research on modeling of passenger demand, service supply, and route choice models. Chapter 3 discusses the design and implementation of the model. It describes in detail the inputs and outputs, assumptions and algorithms of the model. Chapter 4 presents the model results for the London Underground. It then compares the results to outputs from models currently in use by LU and explores alternative assumptions to those made in the model. Chapter 5 summarizes the results of this research, draws conclusions and highlights areas where further work is needed to improve the model, and possible future applications of this model.

2 Literature Review

This chapter will first review prior attempts to solve the problem of identifying the train(s) a passenger took to travel from his origin to destination in section 2.1. It will then discuss literature relating to parts of the problem, specifically: passenger demand estimation (section 2.2), path choice modeling (section 2.3) and assignment of passengers (section 2.4).

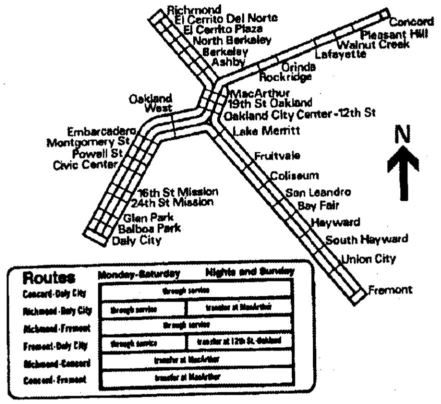
2.1 Prior Works on Train Assignment

To identify the train(s) a passenger took to travel between his origin and destination, a researcher must model the interaction between passenger demand and the transit network. This section will discuss two papers that attempt to model this interaction.

Buneman (1984) modeled the interaction between supply, demand, and transit network using operational data and AFC data together for the first time. Motivated by the need to measure operational performance in terms of the individual passenger, Buneman presents an operational-data based representation of service supply, using the San Francisco Bay Area Rapid Transit (BART) automatic train tracking system. BART has central computer train control, which makes a record of every train action, including each time a train opens or closes its doors. Operational data comes in the form of a list of train actions that detail each arrival and departure of a train at a station in the transit network. These train actions are then converted to train runs, which is essentially a sequence of train actions from train reversal to train reversal. Arrival and departure times at each station are recorded as the time the train door closes. This method has several advantages over the schedule-based method (discussed in detail in section 2.3.2). The main advantage is that it allows the modeler to study actual irregularity instead of simulated irregularity. This leads to deterministic measurement of performance of service in relation to the schedule. Furthermore, operational data-based modeling allows for the day-to-day measurement of impact of service on passengers.

To represent demand, Buneman uses AFC data from BART's stored-fare magnetically encoded ticket system, in which fare is determined at the exit gate by referencing the entry station. For this reason, passenger demand data contains both entry and exit station data. Time of exit is recorded at the two-minute level, but time a passenger entered the system is unknown. With this data, Buneman is able to create OD matrices at the two-minute time period level.

The transit network representation includes stations, lines and transfer requirements. In the BART network, even though some OD pairs are served by multiple routes, the routes consist of parallel and similar services, differing little in travel time, and having many transfer opportunities between the lines that served these routes. Furthermore, routes are decided on the platform. Routes are selected by passengers by assuming the shortest and most convenient route from origin to destination which is defined by the departure and arrival times of trains. To simplify the route choice process, Buneman assumes a single transfer station between routes that serve the same OD pair. For example, passengers traveling from Fremont to Concord may transfer at Oakland City Center-12th St, 19th St Oakland, or MacArthur (see figure 2-1). The program assumes that MacArthur is the preferred transfer station when there is a choice. This simple manner of representing the transit network is sufficient because the BART network is fairly simple and has little ambiguity.



Routes	Monday-Saturday	Nights and Sunday
Concord-Daly City	through service	
Richmond-Daly City	through service	transfer at MacArthur
Richmond-Fremont	through service	
Fremont-Daly City	through service	transfer at 12th St.-Oakland
Richmond-Concord	transfer at MacArthur	
Concord-Fremont	transfer at MacArthur	

Figure 2-1 BART System Map (Buneman, 1984)

Buneman assigns passengers to trains in a "Passenger Flow Model (PFM)." This model computes performance metrics by matching each passenger (from AFC data) with an actual train run (operational data). This matching is done using a deterministic reverse-time simulation, which involves following a passenger's trip backwards from exit to entry on the system. This reverse simulation enables the modeler to overcome the obstacle of not knowing the passenger's entry time. Buneman's process is as follows:

1. Start with a passenger at the exit station at the exit time.

2. Load the passenger onto the last train that arrived at that station before the exit time. The passenger's arrival time at the exit station is recorded as the time this train opened its doors. The selected train in part decides the route the passenger chose; it eliminates any routes that this train does not belong to.

3. Follow this train to the previous station on this passenger's route. If transfer is required in the passenger's route, the passenger will be unloaded at an interchange station. Otherwise, the passenger will be unloaded at the entry station. In either case, the passenger's departure time at this station is recorded as the time this train closed its doors.

4. If a passenger was unloaded at a transfer station, the passenger is loaded onto the last train that arrived at this transfer station before the passenger's departure time. This process is continued until the passenger arrives at the entry station.

5. Record this passenger's trip and move onto the next passenger. After all passengers are assigned to trains, performance metrics relating to passenger experience can be calculated.

Buneman concludes that his methodology for assigning passengers to trains is successful: all passengers were assigned to trains within a reasonable computation time. He is able to compute the complete load on actual train runs. He validates the results by walking through trains and counting passengers, and the accuracy was "found to be very high." This indicates that the reverse-time simulation in which passengers are assigned to the last train that arrived at their exit station works fairly well. Route choice also appears to be a non-issue because of the simplicity of the BART network. Had AFC transaction times at origin stations been available, Buneman would likely have opted for forward-time operational data - based assignment (Frumin, 2010).

Kusakabe et al (2009) creates a model for a Japanese railway company that represents service with scheduled train timetable data and smartcard data to represent passenger demand. The authors use Dijkstra's Algorithm to build the transit network and generate possible paths for each OD pair. The smartcard data reports the entry and exit times and stations for each passenger trip. But since the smartcard penetration rate is about 10%, only a small portion of the passenger demand is represented. Furthermore, passengers with travel time greater than 20 minutes are not considered to save computation time. The timetable data reports each train's departure station and time, arrival stations and times, and train identification number, in the form of train links. The timetables contain express, skip stop and local trains and thus allow for multiple feasible routes between many OD pairs, recognizing transfer possibilities. This schedule based assignment model is typical for services with low frequency. In this network, trains are scheduled to depart within 20 minutes of each other. While this model does not account for delays and other departures from the train schedule, delays to train service are not typical in this Japanese network.

Kusakabe models the interaction between demand and the transit network through a temporal and spatial matching process: passengers can take any train that departs after their entry time at the origin station, but only if that train arrives before the passenger's exit time at the destination station. If passengers have multiple feasible train itineraries, the following rules are applied to assign the passenger:

1. The passenger chooses the train itinerary that has the minimum access, interchange, waiting and egress times.

2. The passenger chooses the train itinerary that has the minimum number of transfers and satisfies the first condition.

3. If there is still more than one itinerary that satisfies the first two conditions then a train itinerary will be chosen at random.

Kusakabe concludes that his methodology for assigning passengers to scheduled trains is successful because the model was able to assign the majority of smartcard passengers to scheduled trains. However, there was a small percent (1.3%) of smartcard passengers who could not be assigned to trains. Detailed analysis of this error suggested that delays and extra trains that are not recorded in the scheduled train timetable data might affect his model's ability to assign passengers to trains. To eliminate this error, Kusakabe recommends the use of operational data to represent service. Kusakabe is able to load smartcard passengers on each scheduled train, but is unable to estimate the complete load on each train and validate the results because he assigns less than 10% of passengers travelling on the rail network. He recommends scaling the smartcard passenger load on each train by some expansion factor and then comparing these results with load weigh data. Finally, he states that using equal probability to choose the route and train itinerary for a passenger (rule 3) may not be appropriate because "each passenger seems to have their own preference. " In other words, this rule does not capture the different utilities possessed by each route and train itinerary, which imply different choice probabilities.

2.2 Passenger Demand Estimation

In order to identify the trains(s) a passenger takes to travel from his origin to destination, the origin, destination, and entry and exit time must be known for every passenger in the transit system. In the context of the London Underground, this level of detail is possible only with automatic fare collection data and OD matrix estimation. This section discusses prior research that developed the methodology for using AFC data to estimate OD matrices.

AFC data offers an inexpensive and efficient way to estimate passenger OD flows on a transit network, whether the transit system requires only entry fare control (typical of systems with flat fares) or uses entry and exit fare control (typical of systems with distance-based or zonal fares). For systems with only entry fare control, like the Metropolitan Transportation Authority New York City Subway, Barry et al (2002) describes a method called trip chaining that is used to infer a passenger's destination by extracting the sequence of transactions the passenger made throughout a day. There are two primary assumptions that allow destination inference. "The first is that a high percentage of riders return to the destination station of their previous trip to begin their next trip. The second is that a high percentage of riders end their last trip of the day at the station where they began their first trip of the day." The resulting OD flows were validated by travel diary information; it confirmed that both assumptions are correct for 90% of passengers. The results were further validated by comparing inferred destination totals by station with station exit counts. Once passenger trips are created by finding a destination for all entry transactions that can have their destinations inferred, these OD flows can be aggregated into an OD matrix, which includes the total number of passengers travelling between each origin-destination pair.

For systems that require both entry and exit fare control, such as the London Underground, an OD matrix can be created directly from the smartcard data by combining entry and exit transactions into a passenger trip, and aggregating each passenger trip by origin and destination. Furthermore, because of the timestamps associated with entry and exit transactions, smartcard data allows for a time period level representation of the OD matrix. In other words, a separate OD matrix can be created for any time period of interest that is sufficiently long given the granularity of the data available. A method to do is discussed below.

One difficulty with smartcard data is that for some transit networks, the data might not represent all passengers using the network. These transit networks may offer a range of fare media besides the smartcard. Another difficulty is that fare collection may not be complete when certain stations in a transit network are not fully gated. Both issues arise in the case of the London Underground, where at least 25% of passengers use fare media besides the Oyster Card. Furthermore, there are at least 24 (out of 270) stations in the LU network that are not fully gated (Gami, 2010). Gordillo (2006) presents a methodology for OD estimation that accounts for the unrepresented (those not using smartcards) passengers and missing AFC data from not fully gated stations for London Underground. Gordillo's methodology is as follows:

1. Create a seed OD matrix from complete smartcard data.

2. Estimate total entry and exits counts at gates at station. This total should include entries and exits for passengers using all types of automatically collected fare media offered by the transit system. For stations that are not fully gated, adjusted manual counts of entries and exits are necessary to compute the total. Adjusted manual counts are manual counts scaled so that when added to the total entry and exit gate counts, they yield the best possible approximation to the entry and exit estimates from survey data.

3. Estimate expansion factors to control for bias in the seed matrix and ensure that the travel patterns are representative of the travel patterns of all passengers using the transit system once the dataset is expanded. Apply the expansion factors to every trip. This creates a singly constrained OD matrix, where the entry totals of the expanded OD matrix matches the total of all passenger entries.

4. Run an iterative proportional fitting (IPF) process, which is a row-column balancing process that transforms the resulting matrix from the previous step into a doubly constrained OD matrix, where the resulting OD matrix entry and exit totals matches the total of all passenger entries and exits.

Chan (2007) expands this work by estimating the OD matrix at the time period level taking advantage of the timestamp associated with smartcard transactions. Chan modifies the process outlined above to create an OD matrix for time period P as follows:

1. Create a seed matrix from smartcard data by allowing only transactions within time period P.
2. Estimate total entries and exits at each station within time period P.
3. Estimate exit proportions to estimate the number of exits that correspond to entries during time period P. Exit proportions are measured for each time interval T at each station A. For example, if P is a three-hour period during the AM Peak (7-10 AM), T would ideally be one-hour intervals within P and extending beyond P (i.e. 7-8 AM, 8-9 AM, 9-10 AM, 10-11 AM, and so on). The exit proportions are a ratio of all completely documented journeys (trips from AFC data with known origins and destinations) that start during time period P conditional on ending at station A during time interval T to all completely documented journeys that end at station A during time interval T regardless of start time. Scale the exit totals for time period P with exit proportions.
4. Estimate expansion factors for each station during time period P and apply expansion factors to the seed matrix to create a singly constrained OD matrix.
5. Run the IPF process that transforms the resulting matrix from the previous step into a doubly constrained OD matrix that matches both station entry and scaled exit totals.

This research suggests that OD estimation at the time period level can be based primarily on AFC data, supplemented with count and survey data when necessary (i.e. in the case where stations are not fully gated). Both Gordillo and Chan demonstrate that using AFC data provides a cost effective, easy to update, and accurate basis for estimating the demand on a rail transit network.

2.3 Path Choice Modeling

Modeling path choice is necessary to identify and characterize the travel options available for each origin destination pair in a transit system. This process illustrates how service lines are connected, enumerates the many ways these lines of service can be used to get from an origin to a destination, and differentiates between these possibilities in terms of convenience, or utility. Wilson (2004) states, "path choices cumulatively determine the spatial distribution of the passenger flow in a network." The first subsection describes select methods for enumerating the set of feasible paths for an individual traveler. The second subsection describes select methods for ascribing routes to a passenger, or groups of passengers.

2.3.1 Path Choice Set Generation

K-shortest path

Guo (2008) outlines the most common methods for path choice generation. First, the k-shortest path algorithm is a deterministic method that generates the first *k*-shortest paths for each OD pair by successively removing a link from the shortest path and finding the next shortest path. Guo states that the drawbacks of this method are that it relies heavily on paths revealed by survey and unrevealed paths may be underrepresented and that it is not flexible and computationally expensive. An alteration of the k-shortest path algorithm is the link elimination method, in which a link from the current shortest path is eliminated at each iteration to generate the next shortest path. This improves upon the k-shortest path method by being more computationally efficient.

Simulation

Prashker and Bekhor (2004) describe the use of simulation for path generation. This method does a shortest path search while drawing a sample of link attributes from assumed distributions. Guo (2008) adds that while this method results in a good coverage of observed paths, it might require a high number of iterations, and the assumed distributions that this method samples from may often be unjustified.

Labeling

Guo (2008) uses a labeling approach to a path choice set by systematically changing the shortest-route criterion. The resulting choice set consists of all labeled shortest paths that are each optimal for a specific label from a given label set. Guo chooses this method because it has been proven to perform well in generating reasonable routes with good coverage of observed paths, and requires much less computation time than the alternatives.

2.3.2 Path Choice Selection

Frequency-Based Model

In the frequency-based assignment model, supply is represented as lines of service with aggregate attributes such as average line frequency. This type of model allows for the calculation of average on-board loads and average performance, and is typically applied to high frequency networks because passengers tend to arrive randomly.

Spiess and Florian (1989) present a transit assignment model that uses the frequency-based model to represent transit service. The transit network is represented by a set of nodes, and a set of lines that are each defined as a sequence of nodes at which passengers may board and alight. At each node, the frequency and average headway of each individual line and all lines combined are known. In this model, the waiting time for each passenger at a particular node is computed by assuming that passengers wait on average half of the headway at that node and that frequencies are combined linearly. Average on-board loads for each line are calculated by assigning passengers to lines based on the probability that the line will be boarded, which is the ratio of its frequency divided by the combined frequency of all lines at that node. While these computations are satisfactory for aggregate measures of system performance, Nuzzolo (2004) states when there are time dependent characteristics of supply that need to be represented and analysis of load on each vehicle is necessary (i.e. timetable design or evaluation of low frequency services), this type of model is not satisfactory.

Schedule-Based Model

In the schedule-based assignment model, individual vehicles are represented as elements of lines of service which allows the measurement of performance for each vehicle, as well as the aggregate measurement of performance of a line of service. These models are typically applied to low frequency transportation networks when passengers are commonly observed to have knowledge of the schedule. One should note that use of scheduled vehicle trips as a representation of service on a transit line implies that the modeled service is regular. Therefore, if one desires to introduce service irregularity into the schedule-based model, this irregularity must be simulated, implicitly or explicitly. While this schedule-based service model allows for time dependent characteristics of supply to be modeled and transit service to be studied at a disaggregate level, it falls short when the realized service varies greatly from the scheduled service and the passenger demand data has actual entry or exit times. Furthermore, it is not appropriate for use in high frequency services.

Nuzzolo (2004) presents a model that uses the schedule-based approach to represent the supply of service. In this model, transit services are represented by a space-time “diachronic” graph that consists of three sub-graphs: a service sub-graph, a demand sub-graph and an access/egress sub-graph. This diachronic graph is pictured in figure 2-2. The service sub-graph consists of nodes representing the scheduled train arrival and departure times at stops, and links representing the scheduled run from one stop to another and the dwell time of the train at a given stop. The demand sub-graph consists of nodes simulating passenger space-time trip characteristics, according to passenger departure and arrival times, and space, according to the physical network. The access/egress sub-graph is the connection between the service and demand sub-graphs with boarding and alighting links. Passengers are then assigned to specific train runs from different transit lines based on the departure times the runs, instead of being assigned to transit lines based on the frequency of the line.

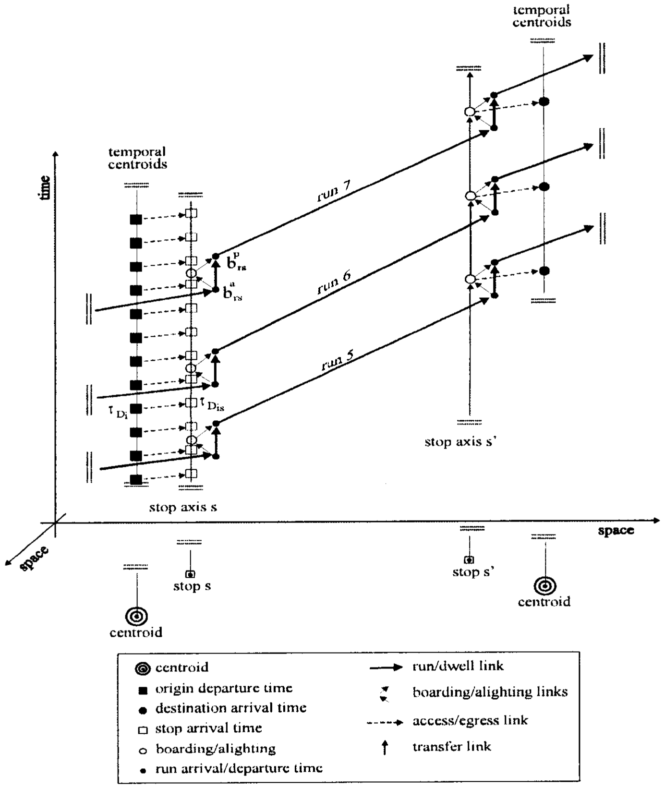


Figure 2-2 Diachronic Graph Representation (Nuzzolo et al, 2001)

Another example of a schedule-based assignment model is Kusakabe et al (2009) discussed earlier in section 2.1

Random Utility Model

In random utility models (RUM) a path is viewed as a choice a passenger faces that is judged based on its objective attributes such as speed, travel time, number of transfers, congestion, etc. These attributes describe the utility, or attractiveness of a path. These paths may contain legs (links) and stations (nodes), and some paths may have common legs and stations. The basic type of random utility models for path-choice modeling is the multinomial logit (MNL). This model allows for the probabilistic choice between multiple paths. This is necessary because the utility of each path to each user cannot be measured directly, and there may be some attributes that cannot be modeled. Ben Akiva and Lerman (1985) describe the utility of path *j* for a particular user as a random variable U_j, shown in equation 2-1.

$$U_j = V_j + \varepsilon_j$$

Equation 2-1

U_j has a deterministic component, V_j, and a random component ε_j. V_j contains the deterministic attributes for path *j* and weights for each attribute. For a path *j* with *k* attributes, x_{jk} represents the attribute itself and β_k represents the weight for that attribute. This is described in equation 2-2.

$$V_j = \sum_k \beta_k \, x_{jk}$$

Equation 2-2

The probability of a passenger choosing path *j* out of a set of *J* alternatives depends on ε_j being independent and identically distributed (i.i.d.), and is given by equation 2-3.

$$P(j) = \frac{e^{U_j}}{\sum_{i \in J} e^{U_i}}$$

Equation 2-3

An important assumption in this model is that passengers' preference for a leg is assumed to be constant and independent of the paths containing this leg. In other words, paths are assumed to be independent and identically distributed. However, if the attractiveness of a leg contributes in the same way to the attractiveness of all paths containing that leg, then there will be correlation between the paths that share that leg. This violates the i.i.d. assumption.

Guo (2008) presents a modified MNL model to account for this shortcoming by specifying explicitly the link and node correlation in the path utility function. Fewer possible overlaps between paths in a transit networks makes it a feasible approach. Equation 2-4 is the modified version of equation 2-2.

Equation 2-4

$$V_j = \sum_k \beta_{jk} x_{jk} + f_n(N) + f_l(L)$$

N and *L* are the set of nodes and links that are included in at least two paths. f_n and f_l are functions specified for a particular node or link, and are applied to all nodes and links contained in path *j*. Guo uses this to model path choice in the London Underground network. *N* contained 23 major interchange stations. With this modified utility function, the probability of each alternative path can be determined through equation 2-3 and passengers can be assigned to paths using this probabilistic approach.

3 Model Development

This chapter presents the model developed to assign the passenger origin-destination matrix to train itineraries on the London Underground network. It discusses the input data required, the algorithms, methods, and assumptions underlying the model, and the model output. The first section (3.1) describes the overall structure of the model. Section 3.2 defines each type of input data including its source, uses and issues associated with its use. The four subsequent sections describe the main processes that are represented in this model: passenger demand process, transit service process, transit network process, and transit system process. The final section (3.7) discusses the output from the model.

3.1 Model Structure

This model is designed to represent a transit system by modeling the interaction between three main processes: passenger demand, transit service and the transit network. Figure 3-1 shows how these processes interact with each other and feed into the overall transit system model. Each process starts with raw data and transforms it into a form that can interact with other processes. Finally, the system model takes the formatted passenger demand data, transit service data and transit network data and produces transit system statistics. The processes and the system model are described briefly below.

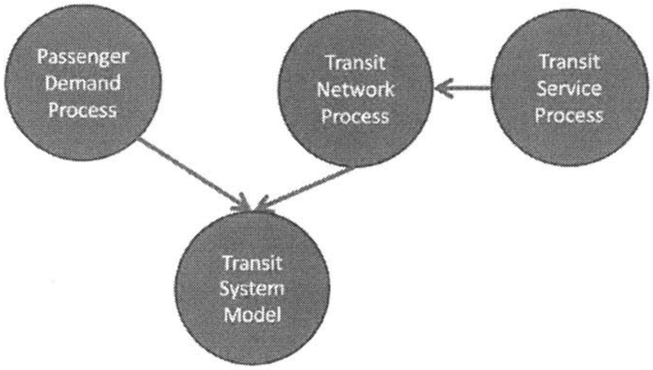


Figure 3-1 Overall model structure

Passenger Demand Process: This process integrates various sources of passenger data to create a list of trips generated at each station each minute with each having an assigned destination. The passenger data include smartcard data, station entry and exit counts, and survey data relating to passenger flow.

Transit Service Process: This process takes as input a list of train events and produces a set of train trips reflecting train movements which represent the service supplied in the network.

Transit Network Process: This process takes as input the train trips (from the Transit Service Process), and network data that defines lines, in terms of stations served and routes between each pair of stations. It combines these two types of data to create a structure that represents the transit network that passengers can traverse. This structure allows the pâssenger choice between routes, and between trains on a route to be represented.

Transit System Model: This model takes as input the results from all three processes and implements algorithms first to generate all possible train itineraries a passenger could have taken to travel between an origin and destination, and then model the passenger selection of route and train itinerary.

Each of these processes will be described in detail in the remainder of this chapter.

3.2 Input

This section identifies and describes in detail all input data required to apply the model. While the model as designed is built for the input files available to the London Underground, the elements of the model that interact directly with the raw data files are isolated so that different data structures can easily be accommodated. The input data includes:

1. Oyster Card data: LU's automatic fare transaction data
2. Automated gate counts: counts of passenger entries and exits at each LU station
3. Manual counts: manual counts of passenger entries and exits at selected LU stations
4. RODS: LU's passenger travel survey
5. Access, Egress and Interchange values: average walk times at each LU station
6. NetMIS data: LU's operational data
7. Route Choice Model: model that describes route choice in the LU network

3.2.1 Oyster Card Data

The first and one of the most critical inputs to this model is automatic fare collection (AFC) data consisting of transactions at the start and end of a passenger's LU journey. For the London Underground, the dominant payment medium is the Oyster card, a contactless smart card which accounts for about 80% of all transactions in the LU network.

The Oyster card, used by passengers riding most Transport for London services including the Underground, Overground, Bus and National Rail services, records transactions for each trip an Oyster passenger makes (Transport for London, 2009b). Because TfL employs a zonal fare system on the Underground, passengers are required to validate their cards both upon entry to and exit from the Underground network. Each pair of linked (through Oyster card ID and transaction time) entry and exit transactions is combined to create an Oyster record that describes a passenger trip from an LU origin station to an LU destination station. Each complete record includes a wealth of information about the ticket type, payment method, and most importantly, trip origin and destination. The key Oyster data for this model are the stations at which the passenger entered and exited the system, and the times at which the passenger passed through the fare gate upon entry to and exit from the Underground network. Many LU stations have multiple fare gates, but the available Oyster data does not include the fare gate information.

For this research, Oyster data from a single day (May 19, 2009) during the AM Peak (7 – 10 AM) on the London Underground network will be used. Table 3-1 lists the relevant fields in the Oyster dataset and table 3-2 provides a sample of Oyster records.

Field Name	Content
PID	Unique encrypted ID for each Oyster card
ENTRYSTN	Name of entry station
ENTRYNLC	National Location Code for entry station
EXITSTN	Name of exit station
EXITNLC	National Location Code for exit station
ENTRYTIME	Time of entry transaction in minutes past midnight
EXITTIME	Time of exit transaction in minutes past midnight

Table 3-1 Oyster Dataset Fields

PID	ENTRYSTN	ENTRYNLC	EXITSTN	EXITNLC	ENTRYTIME	EXITTIME
1	Finsbury Park	580	King's Cross	625	497	507
2	Brixton	778	Victoria	741	585	596
3	Victoria	741	Leicester Square	631	426	438

Table 3-2 Sample Oyster Records

Oyster data has a great many uses and has been the focus of a good deal of research, including its use in the creation of a full OD matrix, as discussed in section 2.2. Another important use is to measure individual passenger travel time on the LU network, and produce travel time distributions for each OD pair, as described in Gordillo (2006), Chan (2007), Uniman (2009) and Frumin (2010). Finally, knowing an individual passenger's travel time as well as his origin and destination time and location allows for analysis of which train in a transit system the passenger might have taken, if data on train trip times is also available. The fact that this level of information is available is crucial to the feasibility of producing a model of this type.

While there are significant advantages to the Oyster data, it also has some limitations as listed below. In the remainder of this section, each limitation will be discussed in detail. Possible solutions to these limitations will be discussed in sections 3.3 through 3.7 as they arise in the model development.

a) Timestamp truncation
b) Clock misalignment at fare gates
c) No route choice information between OD pairs
d) Non-Oyster passengers

a) Timestamp Truncation

The entry and exit timestamps are truncated at the minute level: in other words, seconds are not recorded. In table 3-2, the first passenger is recorded as entering King's Cross station at 497 minutes past midnight, or 7:17. In actuality, he may have entered at any time between 7:17:00 and 7:17:59. TfL plans to remedy this problem in the future, but not within the time frame of this research. This presents a difficulty since desirable temporal precision is lost. Because the results of the model presented in this thesis can be significantly affected by differences in the order of seconds, this is a significant obstacle to overcome. The timestamp 7:17 can mean 7:17:00, or 7:17:59, or any time in between, and this uncertainty can have a potentially significant impact on the results of the model. For this reason, different approaches to overcoming this problem will be tested in the exploration of assumptions in Chapter 4.

b) Clock Misalignment

The clocks at fare gates that record the times for each Oyster transaction are not synchronized and there is strong evidence that for some LU stations, the timestamp is incorrect. The issue stems from the clocks at different fare gates being misaligned, with the degree of misalignment believed to be up to several minutes. Again, TfL plans to remedy this issue in the future, but not within the time frame of this research (Roberts, 2010). Clock misalignment can lead to difficulties in identifying the trains that the passenger might have taken. In order to avoid potential errors associated with this clock misalignment, an approach to overcoming this problem will be discussed later in this chapter.

c) No Route Choice Information

While an OD matrix for the transit network can be produced from Oyster data, this data does not include any information on route choice. When there are multiple routes between a particular OD pair, it is unclear which route the passenger took based simply on the Oyster data. An example of multiple routes between an OD pair can be seen from Oyster data for the third passenger in table 3-1 travelling from Victoria to Leicester Square. From the London Underground map (see figure 3-2), one can see that there are multiple LU routes between Victoria and Leicester Square. A passenger could take the Victoria line to Green Park and then transfer to the Piccadilly line to Leicester Square. Alternately, a passenger could take the District line to Embankment and transfer to the Northern line to Leicester Square. Finally, a passenger could take the Circle line to Embankment and then transfer to the Northern line to Leicester Square. There are still many other possible routes between Victoria and Leicester Square besides these three. For this reason, other data sets must be incorporated into the path choice component of this research. These data sets will identify the possible routes between each OD pair, as well their likelihood.

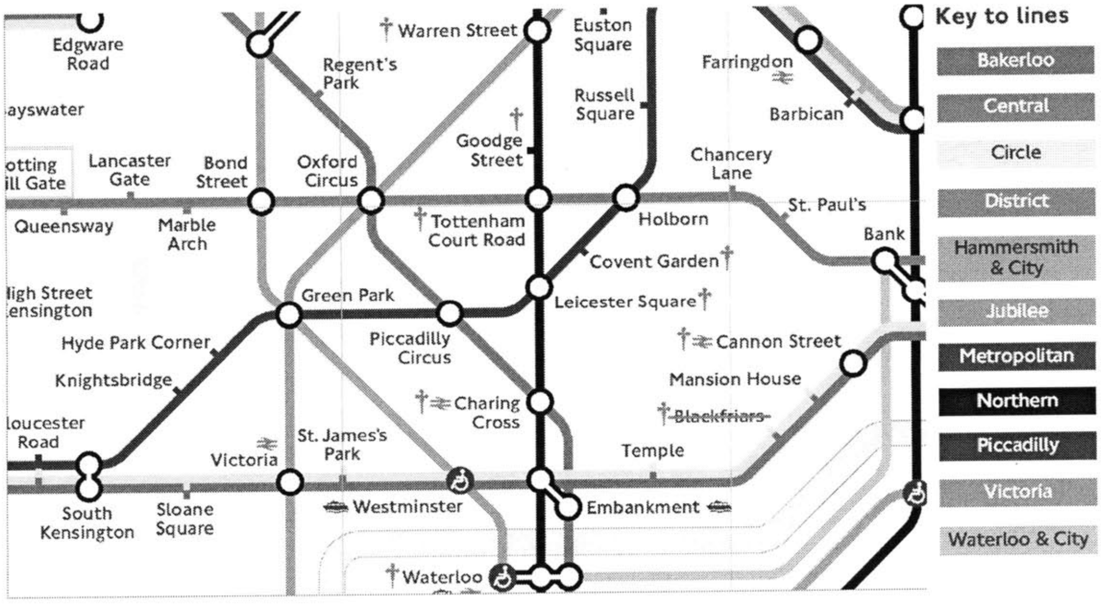


Figure 3-2 London Underground Map

d) Unrepresented Passengers

Oyster data fails to capture all passengers travelling between each OD pair. The most obvious reason for this is that Oyster cards are not the sole fare media for Transport for London services. Passengers can pay for TfL services using Oyster cards, magnetic stripe cards, and paper tickets. Only for Oyster cards can the Oyster OD matrix be directly created because only with Oyster can a passenger's entry be linked to his exit. Non Oyster passengers make up about 20% of all LU passengers. Even among those Oyster passengers, some journeys are not completely recorded. This is because of incomplete transactions—failure to validate at the beginning or end of a journey, often as a result of some stations being not fully gated (NFG). These journeys make up about 7% of all Oyster trips on the LU network on May 19, 2010. In these two situations, there are passengers who are indeed travelling within the transit network, but are unable to be incorporated into the OD matrix based only on Oyster data. To represent the total flow on each OD, other datasets such as passenger flow surveys and entry and exit counts at each gate are used.

3.2.2 Automated gate counts

Automated gate counts are the total entries and exits at each LU station, recorded at fifteen-minute intervals. This data is necessary to expand the Oyster OD matrix to represent all LU customers. Automated gate counts include passengers entering and exiting the network using all types of fare media, including Oyster cards. These counts can be used in combination with Oyster card data to estimate the number of passengers travelling between each OD pair as will be described later in this chapter.

An issue that arises with this data is that some LU stations are not fully gated. This means that passengers can enter and exit the system at these stations without validating their fare media, and therefore there are incomplete counts of the total number of passengers exiting and entering at some stations. To get around this issue, passenger flow surveys are used, as described below.

3.2.3 Manual counts

Manual counts of entries and exits are performed every November at select LU stations that are not fully gated. These counts are performed so that OD estimation at these NFG stations is possible. As of May 19, 2009, there were 20 (out of 270) NFG stations. This number has been significantly reduced since 2007, when there were about 70 NFG stations (Gordillo, 2006). Due to budget constraints, not all NFG stations have manual counts of entries and exits every year. Each year, for the NFG stations not selected, manual counts from previous years are scaled up or down taking into account the change in other gated stations in the same area and zone (Gami, 2010).

These counts are used to supplement the automated gate counts and Oyster card data to estimate the full LU OD matrix.

3.2.4 RODS

One final source of passenger data is a passenger survey. This is a necessary input for this model because it provides valuable information on path choice in the Underground. Specifically, London Underground uses a Rolling Origin Destination Survey (RODS) to estimate OD flows and the flows on alternative routes between each OD pair. This is done by incorporating passenger surveys from a sample of Underground stations over multiple years and expanding results to gate-line counts for each year. On the assigned date, surveys are distributed to randomly selected passengers entering each station. The RODS questionnaire asks about each surveyed passenger's access to the station, trip purpose, route taken within the Underground, times of entry and exit, postal codes of final origin and destination, ticket type and various personal characteristics (Gordillo, 2006).

RODS data is used as an input to many applications that involve passenger demand and behavior in the London Underground. In this research, it is used to supplement the automated gate counts and Oyster OD matrix in developing an OD matrix that represents the total flow of passengers between each LU OD pair.

While RODS data can be extremely useful in revealing the total number of passengers travelling between each OD pair, it is produced infrequently, is expensive to produce, does not capture a significant number of OD pairs that are captured by Oyster data, does not reveal all possible routes between the OD pairs that it does capture, and does not measure day to day or seasonal variation in passenger travel. Chan (2007) concludes that using RODS data as a supplement to Oyster data for applications that involve passenger demand is more appropriate.

3.2.5 Access, Egress, and Interchange Times

This input file describes walk times for every station in the LU network. It is used to link times from Oyster data transactions with times from train movements. For the London Underground, this data comes from the Journey Time Metric. JTM is an Underground service performance measure that emphasizes the customers' experience by evaluating journey time performance, briefly discussed in section 1.4. A passenger's journey is broken down into stages including: access from station entrance to the gate area and then on to platform, interchange between platforms and egress from the platform to the station exit. Each stage has a scheduled value which represents the amount of time a passenger should normally expect to take for the stage. The Access, Egress, and Interchange walk times are measured by manual surveys or through a station simulation model depending on station volume. Access time is measured from the station entrance(s) to the midpoint of platform(s). Interchange time is measured from the midpoint of the arrival platform to the midpoint of the departure platform, assuming the interchange walk starts immediately after the train arrives at the platform. Egress time is measured from the midpoint of the platform(s) to station exit(s), assuming the egress walk starts immediately after the train arrives (Transport for London, 1999).

Manual AEI surveys are conducted at 27 major stations which together account for 46% of all LU passenger demand. These stations are surveyed at least 12 times per time band, for the busiest time bands between 7AM to 7PM on weekdays. (See table 3-3). Data is collected by surveyors walking predefined routes that cover every possible walk route within the station (Transport for London, 1999). PEDS and Legion, pedestrian and station modeling and simulation tools are used to assess the congestion at all stations, including those with AEI surveys. The model incorporates recorded events such as lift and escalator failures and demand fluctuations in estimating the excess AEI times at each station. PEDS and Legion supplement the surveyed AEI data for the 27 major stations and are responsible for the AEI results for the remaining London Underground stations.

Day of the Week	Time of Day	Time Period
Monday-Friday	Early	0530-0700
Monday-Friday	AM Peak	0700-1000
Monday-Friday	Interpeak	1000-1600
Monday-Friday	PM Peak	1600-1900
Monday-Friday	Evening	1900-2200
Monday-Friday	Late evening	2200-0030
Saturday	Morning	0530-1000
Saturday	Midday	1000-1900
Saturday	Evening	1900-2200
Saturday	Late evening	2200-0030
Sunday	Morning	0700-1000
Sunday	Midday	1000-1900
Sunday	Evening	1900-2400

Table 3-3 LU time bands

AEI values are used in JTM to measure the difference between the scheduled and measured walk times for each station. Schedule walk times are defined as 'free flow' timings or the time it would take to walk a route unimpeded. Free flow times take their value from the minimum surveyed walk times for each pedestrian route. AEI values are also used in the model presented in this thesis to help develop distributions for access, egress and interchange walk times for each station. These distributions then provide a basis for identifying which train the passenger could have taken to travel between his origin and destination.

While these values provide a useful basis for developing walk time distributions for each station, it is not clear how accurate they are. For this reason, later in this chapter, the London Underground's average AEI values will be compared against the average AEI values inferred from the model presented in this thesis.

3.2.6 NetMIS Data

Network Management Information System is the primary data source that describes Underground train movements. NetMIS is an event-driven log containing operational data that is derived from the LU signaling system. The signaling system is composed of discrete sections of track called track circuits, varying in length from 50' to 1600', with most being about 500' (about the length of a platform). All rail tracks are included in a track circuit and an Underground line may have about 300 track circuits from end to each in each direction. Some of these coincide (roughly) with platforms. TrackerNet, a track circuit occupancy database, processes and displays track circuit occupancy, getting its data from the various signaling systems. By applying logic to TrackerNet, NetMIS stores in a "train event" each train arrival and departure at stations recorded throughout the day on the LU network. This logic uses the closest track occupancy or unoccupancy event to the station, and applies a fixed temporal offset. For each train event, the observed train arrival and departure times are calculated by adding or subtracting the offset time to track occupancy/unoccupancy records (Rahbee, 2010). Each logged train event includes a unique identification number for each train, the trip number (which increments each time a particular train reverses), the station at which the train arrived (and departed), the calculated arrival and departure times, the line on which the train is travelling, and the direction in which the train is travelling. Table 3-4 lists the relevant data fields from NetMIS data and table 3-5 shows a sample of NetMIS records for a train on the Victoria line on a southbound and northbound trip.

Field Name	Content
TRNEVNT_ID	Train Event identification number
TIMESTAMP	Arrival time of train at station
LINE_ID	LU Line identification number
TRAIN_IDENTIFICATION	Unique identification number for each train
TRIP_NUMBER	Train run number that increments at each train reversal
DIRECTION_CODE	Direction of train run
ACTUAL_DEPARTURE_TIME	Departure of train from station
SUTOR_CODE	Internal three-letter code for station
STATION_NAME	Station name

Table 3-4 NetMIS data fields

TRNEVNT _ID	TIMESTAMP	LINE _ID	TRAIN_ IDENTIFICATION	TRIP_ NUMBER	DIRECTION _CODE	ACTUAL_ DEPARTURE_TIME	SUTOR _CODE	STATION_NAME
105037016	7:00:45 AM	3	1717262	3	1	7:01:28 AM	OXC	Oxford Circus
105037354	7:03:00 AM	3	1717262	3	1	7:03:32 AM	GPK	Green Park
105037664	7:04:59 AM	3	1717262	3	1	7:05:38 AM	VIC	Victoria
105037993	7:07:25 AM	3	1717262	3	1	7:07:48 AM	PIM	Pimlico
105038255	7:09:10 AM	3	1717262	3	1	7:09:36 AM	VUX	Vauxhall
105038675	7:11:50 AM	3	1717262	3	1	7:12:17 AM	STK	Stockwell
105039058	7:14:09 AM	3	1717262	3	1	7:16:54 AM	BRX	Brixton
105039815	7:18:54 AM	3	1717262	4	0	7:19:20 AM	STK	Stockwell
105040217	7:21:15 AM	3	1717262	4	0	7:21:45 AM	VUX	Vauxhall
105040490	7:22:56 AM	3	1717262	4	0	7:23:15 AM	PIM	Pimlico
105040832	7:24:54 AM	3	1717262	4	0	7:25:33 AM	VIC	Victoria
105041141	7:27:04 AM	3	1717262	4	0	7:27:30 AM	GPK	Green Park
105041483	7:29:05 AM	3	1717262	4	0	7:29:31 AM	OXC	Oxford Circus
105041793	7:30:54 AM	3	1717262	4	0	7:31:14 AM	WST	Warren Street
105042027	7:32:29 AM	3	1717262	4	0	7:32:49 AM	EUS	Euston
105042263	7:34:00 AM	3	1717262	4	0	7:35:04 AM	KXX	King's Cross
105042910	7:37:47 AM	3	1717262	4	0	7:38:08 AM	HBY	Highbury & Islington
105043334	7:40:24 AM	3	1717262	4	0	7:40:52 AM	FPK	Finsbury Park
105044094	7:44:46 AM	3	1717262	4	0	7:50:27 AM	SVS	Seven Sisters

Table 3-5 Sample NetMIS data

Figure 3-3 is a graphical representation of NetMIS data for the Victoria line trains travelling north as they arrive at every station on the line. The trains appear to arrive at each station at fairly regular intervals. As depicted, some trains terminate at Seven Sisters because there is access to a train depot from that station and many trips end there.

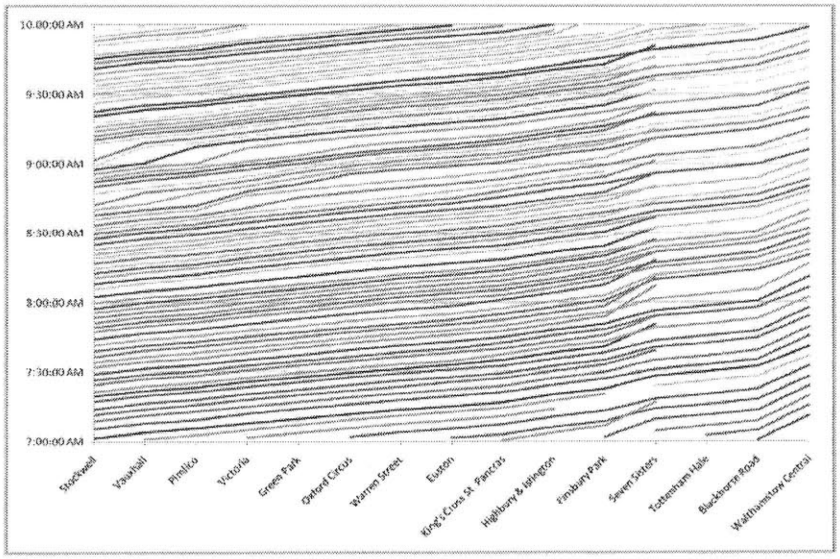


Figure 3-3 Example NetMIS Data for Victoria Line Northbound

NetMIS data has a variety of uses, centered on representing service on the LU network. LU uses NetMIS as input to several models including JTM to measure service delivery. It is also used as an input to LU's train simulation model that analyzes changes to service on the LU network. NetMIS data will be used to represent service available to LU passengers on a specific day. The arrival and departure times calculated in NetMIS will also be used to estimate access, egress and interchange times.

NetMIS data has several important weaknesses listed below which are discussed in the remainder of this section. Possible solutions to these weaknesses will be discussed in sections 3.3 through 3.7 in the context of model development.

a) Missing event records
b) Train identification number reassignment
c) Event record generation at train reversal

a) Missing Event Records

There are significant parts of the network for which NetMIS data is unavailable. Many of these holes in the data also exist in TrackerNet and result from inadequate track signaling equipment. These holes tend to be particularly severe on the sub-surface lines on the LU network, such as the District and Circle lines. This means that there will be some stations or entire branches of LU lines that have little if any

NetMIS data. For example, figure 3-4 represents the Circle line showing that very few trains are recorded over much of this line over a three hour period.

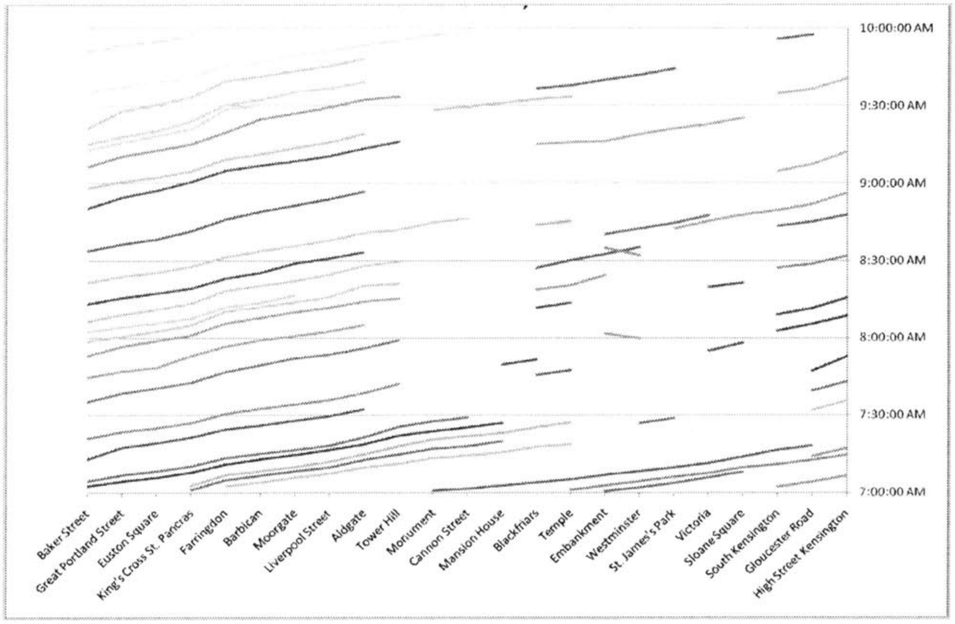


Figure 3-4 Example NetMIS Data for Circle line, Outer Rail

b) Train Identification Number Reassignment

When one track circuit fails to record an occupancy or unoccupancy, TrackerNet sometimes loses track of the train as it passes. When this data is compiled in NetMIS, it appears either as if a train has disappeared for a few stations then reappeared, or that a train disappears at one station, and a new train with a different identification number appears at the next station. This reassignment of train identification number and disappearance of trains is a serious challenge for the model developed in this research. These problems tend to occur heavily on certain lines on the LU network. For example, figure 3-5 representing westbound service on the Piccadilly line over a three hour period shows that most trains are not tracked at certain stations, such as Hyde Park Corner. It is easy to see that as trains pass through this gap in the signaling system, many of them are assigned a different color, which means they are recorded as different trains as they emerge from the gap.

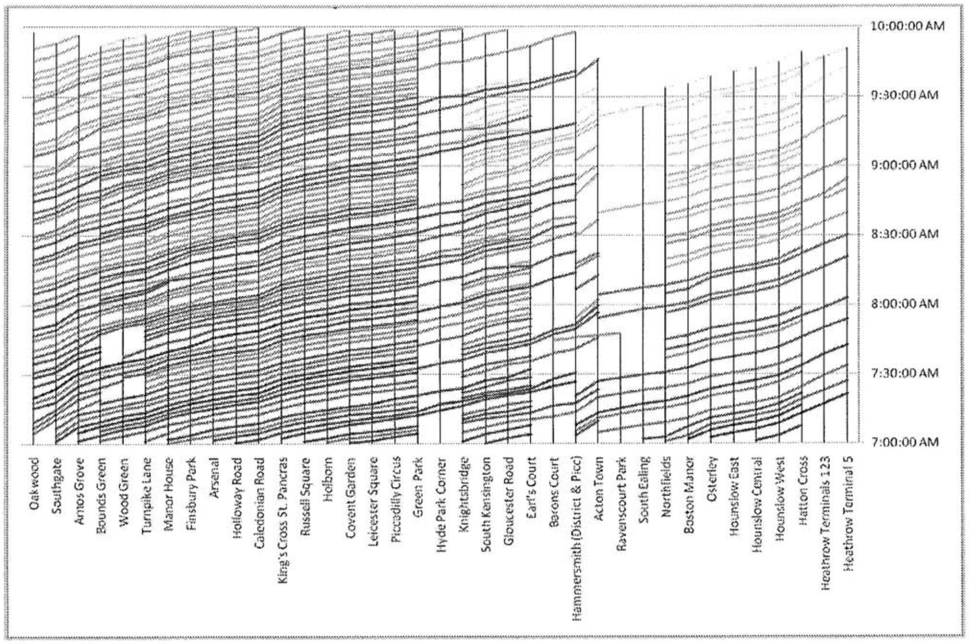


Figure 3-5 Example NetMIS Data for Piccadilly line: Westbound, Heathrow branch

c) Event Record Generation at Train Reversal

When trains reverse at terminals (or elsewhere), an additional train event log is not created. This poses a difficulty when attempting to track complete journeys by trains based on their train identification and trip numbers. For this reason, the starting terminal stations on the Victoria line (figure 3-5) and the Piccadilly line (figure 3-3) are excluded.

To deal with these NetMIS data problems listed above, the model developed in this research requires algorithms and assumptions, and required that the research focus on LU lines that did not have severe NetMIS data problems. These algorithms and the proposed methodology to produce accurate results for selected lines on the LU network (Victoria, Jubilee, and Central) will be discussed later in this chapter.

3.2.7 Route Choice Model

The Route Choice Model describes the transit network from the passenger's perspective. RCM is listed as in input to this model because it is a data file that is read in at the start of the model along with Oyster and NetMIS data. However, it should be noted that the RCM is much more than simple data set. It provides much of the structure to this model. RCM is a model that was created to generate the path choice set for each OD pair on the LU network (Weston, 2009). A path is a sequence of LU lines, connected by interchange stations, which a passenger might take to travel between an origin and a destination station. An LU service line is a concept that represents a set of train tracks (and tunnels) that connect a series of stations and the trains that run on these tracks and serve passengers at these stations. For the most part, though with some notable exceptions, an LU service line does not share its tracks with other service lines and has a main trunk portion. Most lines also have branches, with each branch splitting from the trunk. An LU station might be served either by a single line, or by multiple lines, in which case it would be called an interchange station. Two arbitrary stations can be connected by LU service lines, which in turn are connected by interchange stations. The set of possible combinations of lines and interchange stations between two designated stations is called the path choice set for this OD pair.

The path choice set used in the RCM was generated using a shortest path algorithm that arrives at alternative routes by eliminating different lines for each run, and optimizing for different journey characteristics, including minimum walk (access, egress, interchange) time and minimum number of interchanges. The RCM includes paths specified by RODS. Each route in each choice set includes:

- Start, end and interchange stations
- Number of interchanges
- LU lines that service each leg of the route
- Expected run time for each leg of the route.

In addition to path choice set generation, the RCM also specifies a generalized cost for each possible path between each OD pair. This generalized cost is a way of assessing the utility of a route relative to others serving the same OD pair. The cost function is defined in Equation 3-1 with the coefficients shown being the standard values recommended in TfL's Business Case Development Manual (2009).

Generalized Cost = (Run Time) + 0.7*2(headway) + 2(interchange time)
+ 4(access + egress time) + 3.5(number of interchanges) **Equation 3-1**

Table 3-6 gives an example of the path choice set for travel between Victoria station and Leicester Square.

ID	Entry Station	Exit Station	Total Generalized Cost	Num of Interchanges	First Line	First Direction	First Run Time	Int 1 Station	Int 1 Line	Int 1 Direction	Int 1 Run Time
1	Victoria	Leicester Square	26.4	1	Circle	Inner	5.5	Embankment	Northern	North	2.5
2	Victoria	Leicester Square	29.1	1	Victoria	North	2	Green Park	Piccadilly	East	3.3
3	Victoria	Leicester Square	37.7	1	Victoria	North	10	King's Cross	Piccadilly	West	6.8
4	Victoria	Leicester Square	33.4	1	Victoria	North	6.4	Warren Street	Northern	South	4.3
5	Victoria	Leicester Square	33.4	1	Circle	Outer	4.8	South Kensington	Piccadilly	East	9.9

Table 3-6 Path Choice Set between Victoria and Leicester Square

The Path Choice Set generated by the RCM can be used in a number of planning models that require enumeration of all possible paths. For the model developed in this research, these routes serve as a mold into which itineraries built from trains from NetMIS data must fit.

3.2.8 Summary of Inputs

This section described all inputs to the model to infer the train(s) that a passenger most likely used to travel from his origin to destination. The inputs described above are specific to the London Underground model application and this discussion has focused on their uses and limitation. This sub-section discusses the inputs that are crucial and those that are helpful if a model of this type is to be developed for another transit system.

Passenger Demand: Crucial passenger demand data includes information about each passenger's entry time and station, and exit time and station in as much detail as possible. Passenger demand data should represent all passengers that travel on the transit system. Supplemental information from demand data may include specifics on which entry/exit of a station the passenger used, which may provide information indicating the route the passenger chose from his origin to destination, and fare card validation at intermediate points of a journey, which may indicate the route taken and the times at which a passenger passed through certain points in the transit network.

Service: Crucial service data includes information about each train's arrival and departure time at each station on the network. The data must be recorded so that each train's arrivals and departures can be identified, and the train's actions can be tracked in chronological order. Each trip a train takes should be differentiated from the other trips. Each train must belong to a group or a service line as defined by the transit network data, and must serve stations that are identified by the transit network data. Supplemental information from service data may include the weight of the train upon arrival and departure at each station, which may provide a proxy for load on the train (Frumin, 2010).

Network: Crucial transit network data includes service line definition, which may be the group of trains that serve each line in the network; all stations served by each defined line; and identification of stations that serve multiple lines. For any pair of stations identified, the data must provide a set of routes, each consisting of a different sequence of lines connected at stations that are served by both lines. This set of routes for a pair of stations in the network represents the set of choices a passenger faces when travelling on the network. Supplemental information may include the expected run time necessary for trains to serve each route. Detailed information about each station including its structure, possible routes between each point of interest such as platforms and gates, and the amount of time it takes for a passenger to walk between each pair of points is helpful.

3.3 Passenger Demand Process

In this component of the model, passenger demand data is processed and then expanded to represent all passengers using the transit system. For the London Underground, there are four sources of raw demand data that go into this process: automated gate counts, manual counts, RODS, and Oyster data. While the first three sets of inputs could be used to estimate an OD matrix that represents all passenger demand on the LU network, they do not provide the precise time of entry and exit that the last input, Oyster data, does. For this reason, all four passenger inputs are combined so that every passenger on the LU network has an estimated entry and exit time. The steps of this process are listed in 3.3.1. In section 3.3.2, a process to deal with Oyster timestamp truncation is introduced. Finally, in section 3.3.3, the creation of passenger trips is described.

3.3.1 Process to assign entry and exit times to all passengers

1. **Estimate full OD Matrix:** Combine the four sets of data using the iterative proportional fitting methodology (described in section 2.2) to represent the total passenger flow on the LU transit network. The result of this process is an expanded OD matrix that contains estimated OD movements for all LU passengers. This process assigns entry and exit counts to OD pairs and therefore estimates OD movements for every passenger on the LU network.
2. **Isolate non-Oyster passengers:** The OD matrix for non Oyster passengers is determined by subtracting the Oyster OD matrix from the expanded OD matrix.
3. **Distribute non-Oyster passengers:** For each OD pair in the resulting matrix, non Oyster passengers are uniformly distributed to all Oyster passengers travelling between the same OD pair. This is done to assign entry and exit times to the non Oyster passengers.

This process effectively groups non Oyster passengers with Oyster passengers. It assumes that non Oyster passengers and Oyster passengers have similar temporal travel patterns within the AM Peak period.

3.3.2 Oyster Transaction Time Truncation

The fact that all Oyster entry and exit times are truncated at the minute level (see section 3.2.1) means that the treatment of time of entry and exit needs some care. In terms of journey time, the true value can be anywhere between the following two extremes:

1. The shortest possible travel time would result if the "true value" of entry times is 59 seconds after the recorded transaction time and the "true value" of the exit time is at the recorded transaction time. This could mean that a passenger's travel time could be up to one minute less than the recorded journey time. If the recorded entry time is used in the model, some train itineraries which were not feasible would be included in the feasible set.
2. The longest possible travel time would result if the "true value" of the entry time is as recorded and the "true value" of the exit time is 59 seconds after the recorded transaction time. This could mean that a passenger's travel time could be up to one minute longer than the recorded journey time. If the recorded exit time is used in the model, some train itineraries which were feasible would be excluded from the feasible set.

These two cases illustrate the trade-offs between feasibility and inclusiveness in treatment of transaction time. For the purpose of this research, the first option is followed in which the entry time is increased by 59 seconds and the exit time is as recorded. This implies the shortest possible travel time (the time between entry and exit) given the recorded transaction times, which will minimize the number of feasible train itineraries. This assumption is chosen because the model can deal separately with passengers for whom the assumption was too restrictive. The system model has a mechanism to handle estimated passenger travel times that are too short. However, there is no mechanism that handles estimated passenger travel times that are too long and which can generate erroneous results. In other words, the design of the main system model requires the shortest possible passenger travel time be assumed. The effect of following the second option is explored in chapter 4.

3.3.3 Passenger Trip Generation

Finally, each linked pair of Oyster transactions (representing a trip by an Oyster passenger) is processed to define the origin, destination, (modified) entry time, exit time, and the number of (additional) non Oyster passengers travelling between the same origin and destination that were "assigned" to this passenger. This level of passenger demand data reflects the information necessary for all passengers on the transit network as input to the train assignment process.

To summarize, the adjustments and assumptions made in the Passenger Demand Process are:

1. Entry time is adjusted to the end of the one-minute feasible window, i.e. fifty nine seconds are added to the time of each entry transaction.
2. Exit time remains as recorded.
3. Non Oyster passengers have temporal travel patterns similar to Oyster passengers. Each non Oyster passenger is randomly assigned to an Oyster passenger with the same OD.

3.4 Transit Service Process

This process takes train operational data, specifically from NetMIS as described in section 3.2.6, and converts it so that it can be matched with the transit network data. A NetMIS table for a particular day consists of a list of all recorded train arrival and departure events for that day. As discussed in section 3.2.6, some issues arise because the NetMIS data is not complete. Some of these issues will be addressed in the Transit Service Process, and others will be addressed in the system model.

The Transit Service Process structures and stores the operational data in a hash table, which is a data structure that efficiently maps certain identifiers or "keys" to their associated values. One searches a hash table by looking up a key of interest, and the associated values are returned. This data structure is used because the average computation time for each lookup is independent of the number of elements stored in the table. Hash tables are used when there are frequent searches required and speed is important. This data structure will be used extensively in the model developed in this research.

NetMIS data is stored in a hash table of hash tables. The outermost hash table contains a set of keys that pertain to each line in the LU network. The associated values for each line contain another hash table. This inner hash table contains a set of keys that pertain to each station on that particular line. The associated values for each station contain a list of train events that occurred at that station. This data structure assists in the process of looking up train events. Instead of reading through the entire NetMIS table to find relevant train events (on a line and at a station of interest), the program can quickly find all the stations, and all the associated events at the station of interest. This structure is designed to facilitate the interaction with the other model processes.

The Train Service Process is as follows:

1. Read each event from the NetMIS file.
2. Discard any events that occurred at depots or stations not on the LU network.
3. Create a new Train Event object. Each Train Event object contains a Train Identification number, Trip number, LU line, arrival time and departure time.
4. If the departure time is missing from the event log, assume the departure time is equal to the arrival time.
5. Store the Train Event under the appropriate line and station in the NetMIS hash table.
6. Sort the list of Train Events under the selected line and station by departure time.

The only assumption in the Train Service Process is that a train's departure time from a station is equal to the train's arrival time at the station when the departure time is missing. About 3% of NetMIS event logs have missing departure times.

3.5 Transit Network Process

The Transit Network Process contains information about lines, stations, and routes between each pair of stations. It takes raw data and outputs from other models and creates a structure that enables rapid searching and matching with other aspects of a transit system. The output of the Transit Service Process feeds into this model and provides structured information about lines of service, stations served by each line, and all trains that served those stations. The other critical input to this program is the Route Choice Model (RCM) produced by LU, which enumerates all reasonable routes between each OD pair. Each possible route describes the services a passenger travelling on this OD pair might take to reach the destination. Each route in the RCM is identified by the origin, destination, number of stations at which a passenger changes services (interchanges), all lines of service and the direction of travel, all stations at which interchanges occur, the expected travel time for each leg of the journey, and the generalized cost for the route. The RCM is described in more detail in section 3.2.7.

This program creates a hash table structure from its two inputs. The hash table contains a set of keys that represent each OD pair, and the values associated with each key are the OD object. Each OD object contains information about the origin and destination stations, and the list of routes between the origin and destination. Each route contains information about its origin, destination, the number of interchanges, and a list that contains each leg or segment of this route. A route is segmented by interchanges and each segment corresponds to a line and direction of service. For example, if a route has two interchanges, it will have three legs. Each leg contains information about the start and end stations, the line and direction of service, the expected run time for the leg, and a list of Train Events that occur on that particular line and station for the start and end stations. The lists of Train Events are populated from the results from the Transit Service Model. This program does not take up unnecessary data storage space because it uses pointers to link ODs, Routes, Legs and Stations to Train Events. This way, Train Events referenced by different ODs, Routes, Legs and Stations are not duplicated. The resulting tree-like structure (see figure 3-6), which is the output of the Transit Network Process, enables immediate access to relevant Train Events and allows the model to avoid repetitive searching when interacting with the Passenger Demand Process.

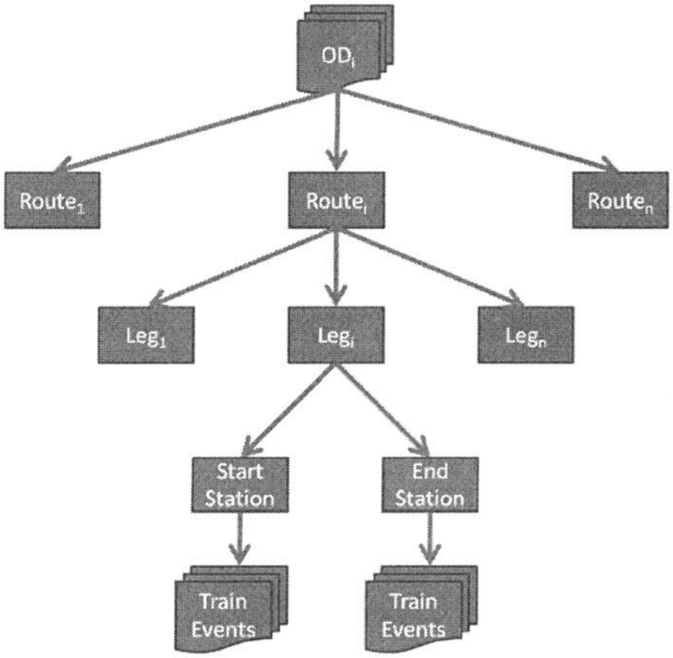


Figure 3-6 Transit Network Process Output

The Transit Network Process is as follows:

1. Read in each possible route from the RCM file.

2. Create a new OD object if it does not already exist and store in the hash table.
3. For each OD, create Route objects and populate the list of routes.
4. For each route, create Leg objects and populate the list of legs.
5. For each leg, store the start and end station.
6. For each station, search the results from the Transit Service Model for the relevant Train Events and populate the list of Train Events.
7. If the Transit Service Model produces no Train Events for a particular line and station, the list of Train Events will be empty.

3.6 Transit System Model: Generation of Itineraries

The Transit System Model is more complex than the other processes described above. It takes the results from the Passenger Demand Process which represents the total passenger demand; the results from the Transit Network Process, which is a data structure that represents the transit network and service; and other data files. The model then processes the data under a series of assumptions, and ultimately produces a set of statistics on the transit system.

There are two main components in the Transit System Model. The first component is the Generation of Itineraries. This component matches the output from the Passenger Demand Model with transit network and service data from the Transit Network Model. The underlying idea is that train itineraries, or sequences of trains that run between a particular origin and destination stations, can be generated when travel time from a passenger's trip data is applied to the transit network data. In other words, given a passenger's origin and destination, and departure and arrival times, it should be possible not only to identify which service lines the passenger may have chosen, but also which trains on those lines the passenger may have taken. A train itinerary is a chronological sequence of trains that a passenger may take from his origin to his destination, including connections at interchange stations. The Generation of Itineraries component of the Transit System Model produces a list of feasible itineraries for each passenger.

The second component is the "Selection of Itinerary." This component takes as input the itineraries generated in the preceding step and applies a series of steps to identify the most probable itinerary taken by a passenger. These steps use data describing walk times in stations. The final output is a single train-level itinerary which, when combined with all the other selected train itineraries, produces a wealth of information that can be used to produce statistics on the transit system. This component will be discussed in section 3.7.

To generate a set of feasible itineraries for each passenger, this component matches passenger demand data with transit network data. The program processes each Oyster passenger record (possibly grouped with one or more non-Oyster passengers) from the passenger demand process output, which has four critical pieces of associated information: origin station, destination station, entry time and exit time and then uses that information to match the passenger with the transit network data. The general process of this program is summarized below:

1. Collect the passenger's origin and destination and find the matching OD pair in the data structure from the Transit Network Model.
2. Generate train itineraries for each route pertaining to that OD pair by retrieving Train Event data from the tree-structure, constrained by the travel time associated with the passenger.
3. Test these itineraries to ensure that they are internally consistent.

This process is described in detail in the following subsections. The second and third steps differ when dealing with complete or incomplete NetMIS data or the existence of interchanges on the routes of interest. The following subsections describe itinerary generation algorithms for the following scenarios: complete data with no interchanges, complete data with interchanges, incomplete data with no interchanges, and incomplete data with interchanges.

3.6.1 Generating Non-Interchange Itineraries with Complete Data

Itineraries are generated for a passenger travelling on a route that has no interchanges by applying temporal and spatial constraints to the candidate Train Events. Train Events from the entry and exit stations are retrieved from the transit network data structure. The number of Train Events retrieved is constrained temporally by the entry and exit times: all feasible Train Events at the entry (exit) station must have departure (arrival) times after (before) the passenger's entry (exit) time. Once feasible Train Events are populated, Train Events at the entry station are linked to Train Events at the exit station by matching Train Identification numbers assigned to each Train Event. If Train Events with matching train identification numbers and the entry station Train Event occurs before the exit station Train Event, then they both are events for the same train, which can be linked to create an itinerary.

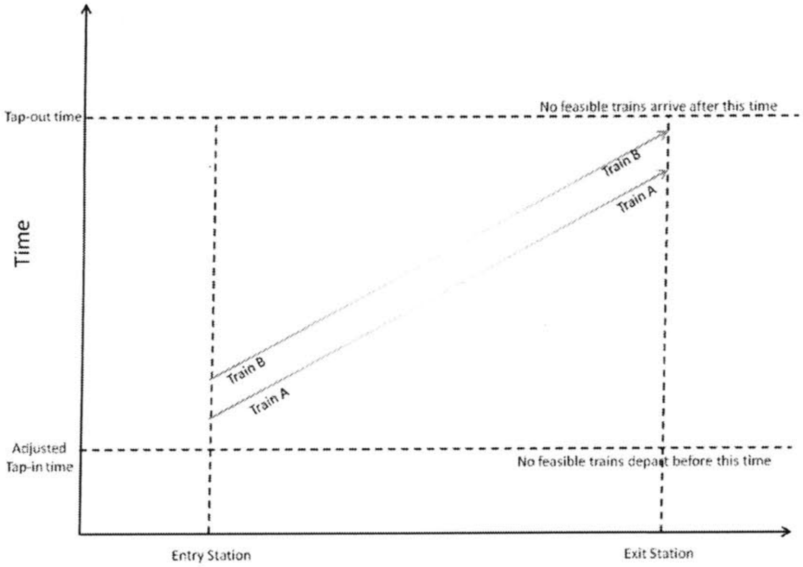


Figure 3-7 Generation of Itineraries: Complete Data with No Interchange

In figure 3-7, two Train Events at the entry station share train identification numbers with Train Events at the exit station and fall within the passenger's entry and exit time constraints. These linked pairs of Train Events are path segments labeled "Train A" and "Train B", which constitute the two feasible itineraries for this passenger.

The following is an example of generating itineraries for a passenger travelling between an OD pair with one possible route with no interchanges on a line with complete data. In table 3-7, an Oyster passenger enters at Brixton at 8:45 (adjusted to 8:46 based on the timestamp truncation adjustment described in section 3.3) and exits at Victoria at 8:59. Table 3-8 shows the two feasible itineraries generated from the transit network data. Time is displayed in minutes past midnight. Each pair of Train Events is linked to form a train itinerary:

- The first itinerary departs Brixton just before 8:48 and arrives at Victoria just before 8:57.
- The second train itinerary departs from Brixton at 8:49 and arrives at Victoria just before 8:59.

Figure 3-8 shows the Oyster passenger's entry and exit time in relation to the two generated itineraries.

PID	ENTRYSTN	EXITSTN	ENTRYTIME	EXITTIME
522213295	Brixton	Victoria	525	539

Table 3-7 Sample Oyster Record: Complete Data with No Interchange

Train	Dep Station1	Dep Train Run1	Dep Train Time1	Arr Station1	Arr Train Run1	Arr Train Time1	Run Time1
Train 1	Brixton	17202313	527.57	Victoria	17202313	536.77	9.20
Train 2	Brixton	17204291	529.40	Victoria	17204291	538.72	9.32

Table 3-8 Sample Generated Itineraries: Complete Data with No Interchange

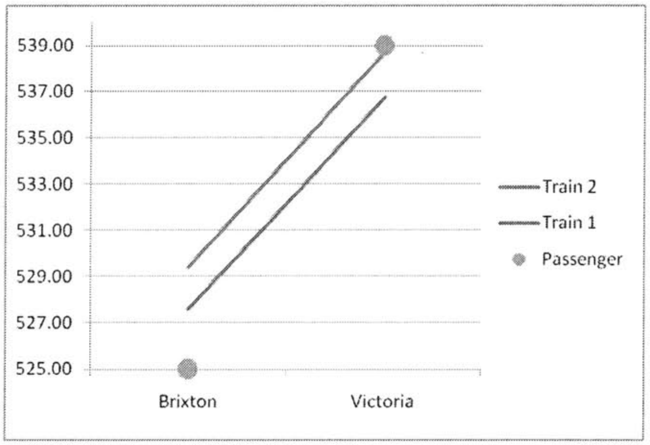


Figure 3-8 Sample Oyster Record with Generated Itineraries: Complete Data with No Interchange

In summary, the algorithm for generating itineraries for routes with complete data and no interchanges is as follows:

1. Identify all Train Events at the entry and exit stations between the passenger's entry and exit time.
2. Link Train Events at the entry and exit stations by matching the train identification numbers and making sure that the entry station Train Event departure time is before the exit station Train Event arrival time.
3. Each pair of linked Train Events is a path segment.
4. Because there are no interchanges, each path segment is also an itinerary.

Using the algorithm described above, figure 3-9 below shows the distribution of passengers by number of feasible itineraries. The majority of passengers traveling on routes with complete data and no interchanges have one feasible itinerary. In other words, for these passengers, there is a unique feasible train itinerary and no selection is required. However, for the remaining 45% of passengers, further selection is required as is discussed in section 3.7.

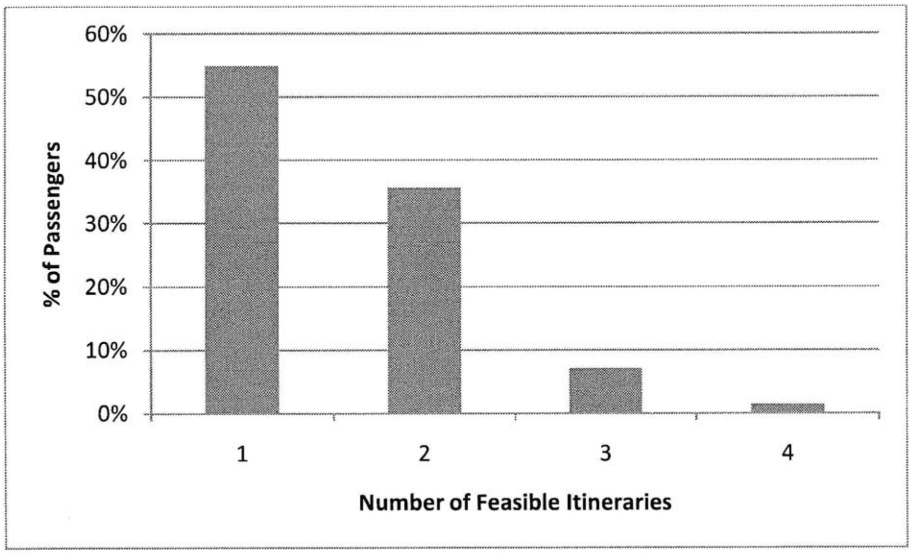


Figure 3-9 Distribution of Passengers by Number of Feasible Itineraries: Complete Data with No Interchange

3.6.2 Generating Interchange Itineraries with Complete Data

Itineraries are generated for a passenger travelling on a multi-leg route with one (or more) interchange using times to constrain feasible path segments on adjacent legs. Train Events from the entry station, exit station and interchange station(s) are extracted from the transit network data structure. The passenger's route shown in Figure 3-10 has one interchange, and therefore two legs. The Train Events retrieved at each station are constrained by the entry and exit times, as in the previous algorithm.

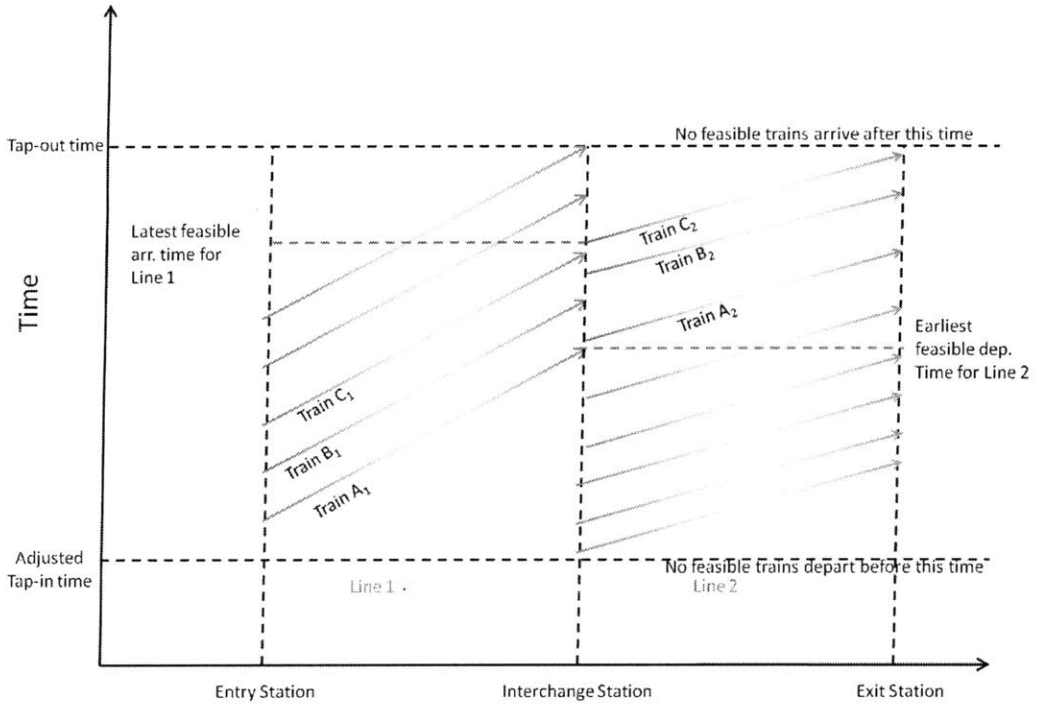


Figure 3-10 Generation of Itineraries: Complete Data with Interchange

Train Events are then linked for each leg of the route. Train Events at the entry station are linked to Train Events at the (first) interchange station by matching the Train Identification numbers. If the entry station Train Event occurs before the Train Event at the interchange station, this represents a possible itinerary leg. Thus path segments are generated for the first leg of the route. The same process is applied to every other leg of the route. For example, in figure 3-10, the path segments will be generated for the second leg by linking Train Events that depart from the interchange station and arrive at the exit station. These path segments for each leg are generated independently of every other leg.

Next, infeasible path segments must be eliminated. It is clear from figure 3-10 that some generated path segments will be infeasible. For example, path segments on the first leg that arrive very close to the passenger's exit time are infeasible because there are no connecting path segments on the second leg that would satisfy the passenger's exit time constraint. Similarly, path segments from the second leg that depart very close to the passenger's entry time are infeasible because there are no connecting path segments on the first leg that would satisfy the passenger's entry time constraint. Thus, the range of feasible arrival and departure times for each path segment on each leg can be used to constrain the number of feasible path segments on adjacent legs. In figure 3-10, this doubly constraining relationship between the two legs is shown by dotted lines.

- A feasible path segment on the second leg must have a departure time after the arrival time of the earliest feasible path segment on the first leg.
- A path segment on the first leg must arrive at the interchange station before the departure time of the latest path segment on the second leg.

Finally, the remaining feasible path segments for each leg are synthesized into itineraries with each itinerary contains one path segment for each leg of the route. There is an itinerary created for every feasible combination of path segments. In figure 3-10, six feasible itineraries are generated: Train A_1 + Train A_2, Train A_1 + Train B_2, Train A_1 + Train C_2, Train B_1 + Train B_2, Train B_1 + Train C_2, and Train C_1 + Train C_2.

The following is an example of generating itineraries for a passenger travelling between an OD pair with a route that involves one interchange with both legs on lines with complete data. In table 3-9, an Oyster passenger enters at Victoria at 8:36 (adjusted to 8:37 based on the timestamp truncation adjustment described in section 3.3) and exits at Leicester Square at 8:48. In this example, only one route between Victoria and Leicester Square will be considered: the Victoria line from Victoria to Green Park for the first leg and the Piccadilly line from Green Park to Leicester Square for the second leg. Table 3-10 shows all feasible path segments for the first leg, generated from transit network data between Victoria and Green Park, between the passenger's entry and exit times. Table 3-11 shows all feasible path segments for the second leg, generated from transit network data between Green Park and Leicester Square, between the passenger's entry and exit times. Figure 3-11 shows the Oyster passenger's entry and exit time in relation to the three generated itineraries. The three generated itineraries are: Train 1 Leg 1 + Train 1 Leg 2, Train 1 Leg 1 + Train 2 Leg 2, and Train 2 Leg 1 + Train 2 Leg 2. Train 3 Leg 1 is eliminated as infeasible.

PID	ENTRYSTN	EXITSTN	ENTRYTIME	EXITTIME
513236301	Victoria	Leicester Square	576	588

Table 3-9 Sample Oyster Record: Complete Data with Interchange

Train	Dep Station1	Dep Train Run1	Dep Train Time1	Arr Station1	Arr Train Run1	Arr Train Time1	Run Time1
Train 1 Leg 1	Victoria	17237824	578.50	Green Park	17237824	580.63	2.13
Train 2 Leg 1	Victoria	17224454	580.55	Green Park	17224454	582.63	2.08
Train 3 Leg 1	Victoria	17213644	582.50	Green Park	17213644	584.63	2.13

Table 3-10 Sample Path Segments for First Leg: Complete Data with Interchange

Train	Dep Station1	Dep Train Run2	Dep Train Time2	Arr Station1	Arr Train Run2	Arr Train Time2	Run Time2
Train 1 Leg 2	Green Park	17229550	581.95	Leicester Square	17229550	585.13	3.18
Train 2 Leg 2	Green Park	17229820	583.48	Leicester Square	17229820	587.13	3.65

Table 3-11 Sample Path Segments for Second Leg: Complete Data with Interchange

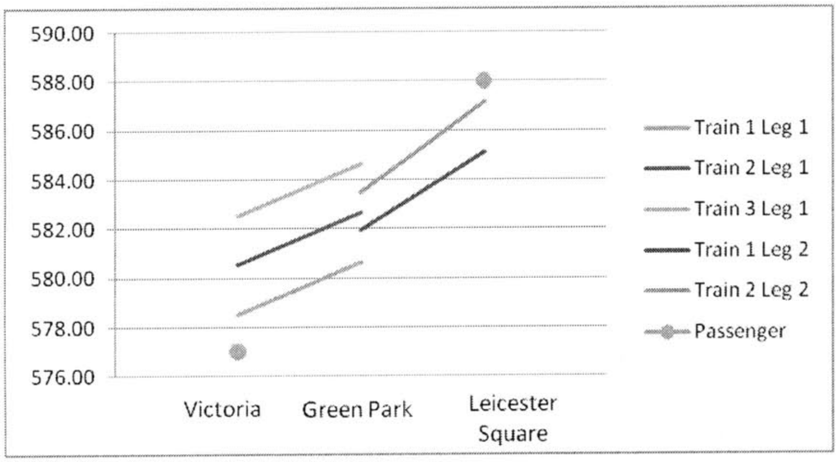


Figure 3-11 Sample Oyster Record with Generated Itineraries: Complete Data with Interchange

In summary, the algorithm for generating itineraries for routes with interchanges and complete data is as follows:

1. Extract all Train Events at the entry, exit, and interchange stations between the passenger's entry and exit time.
2. Link Train Events at adjacent stations (pairs of stations that make up a leg of a route) by matching the train identification numbers and making sure that the departure time of the Train

Event from the station at the start of the leg is before the arrival time of the Train Event from the station at the end of the leg. This should be done for every leg.

3. Each pair of linked Train Events is a path segment.
4. Eliminate all infeasible path segments by using the earliest arrival time or latest departure time of path segments from adjacent legs on the route.
5. Generate itineraries by joining all feasible combinations of path segments.

Using the algorithm described above, figure 3-12 shows the distribution of passengers by the number of feasible itineraries. These passengers travel on routes with one (or more) interchanges and complete data. The highest percent of passengers have one feasible itinerary, though 68% passengers have more than one feasible itinerary. This spread distribution is a result of the duplication of path segments in different itineraries, as seen in figure 3-11. Of course, the number of feasible itineraries is over stated here because it assumes that interchanges can be made with zero elapsed time.

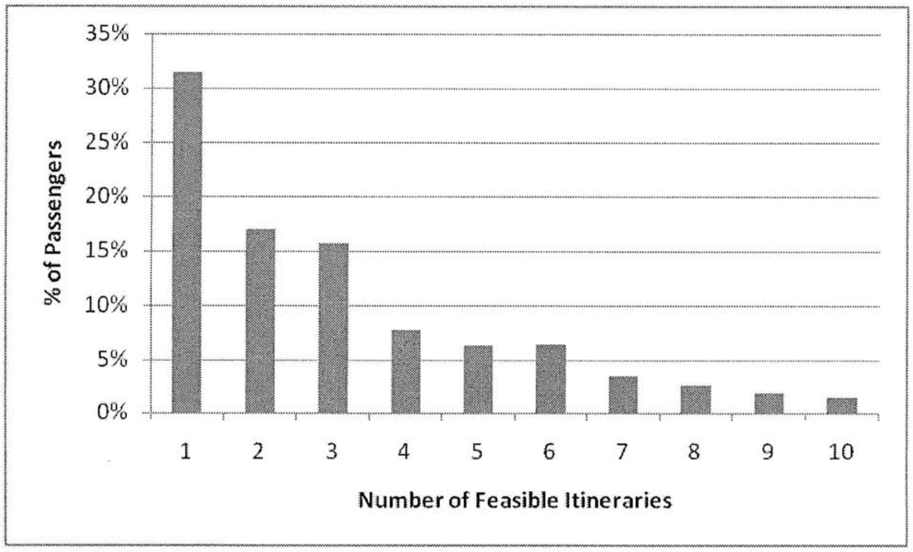


Figure 3-12 Distribution of Passengers by Number of Feasible Itineraries: Complete Data with Interchange

3.6.3 Generating Itineraries without Interchanges and with Incomplete Data

The itinerary generating process includes mechanisms to ensure that at least one itinerary is generated for every trip. These mechanisms are in place to deal with deficiencies in the input NetMIS data. When the input data is improved in the future, these mechanisms should become unnecessary.

This section deals with the challenge of generating itineraries for passengers in the face of incomplete NetMIS data i.e.: operational data that is missing for some LU lines. Because this model is intended to produce complete statistics on the LU transit system, it is not possible to simply ignore the LU lines of service with incomplete data, or passenger trips that involve these lines. Instead, assumptions must be made to salvage as much information as possible.

Three LU lines of service in NetMIS have consistently complete data: Victoria, Central, and Jubilee lines. For passengers that start and end their journeys on these lines, itineraries can be generated based on the algorithms described in section 3.6.1. However, there are a significant number of passengers that interchange between lines with complete and incomplete data. It is important to attempt to generate itineraries for these passengers so that their travel on the complete lines can be captured and comprehensive statistics for these lines computed. This section presents an algorithm that allows for this model to generate accurate itineraries for non-interchanging passengers when they travel on incomplete NetMIS data lines. While results from these passengers are not relevant to the computation of comprehensive statistics of the LU network, generating itineraries for these passengers is a useful exercise in understanding how to generate itineraries for passengers that interchange between lines with complete and incomplete data.

Occasionally, TrackerNet, the database that feeds NetMIS, momentarily loses a train as it passes through the track circuits. These momentary communications malfunctions between the signaling system and TrackerNet throw off TrackerNet's tracking logic, and TrackerNet will generate a new train identification number. This data is then fed into NetMIS making it appear as though new trains appear at stations after the signaling malfunction, and the previous trains disappear. Figure 3-13 shows how trains get reassigned identification numbers after passing through such a gap. It is clear that the entry station Train Event identified as "Train 1" and exit station Train Event identified as "Train 3" are the same train, but the two train Events that pertain to Train 1 and Train 3 have different identification numbers, and are therefore not linked applying the algorithms for complete NetMIS data.

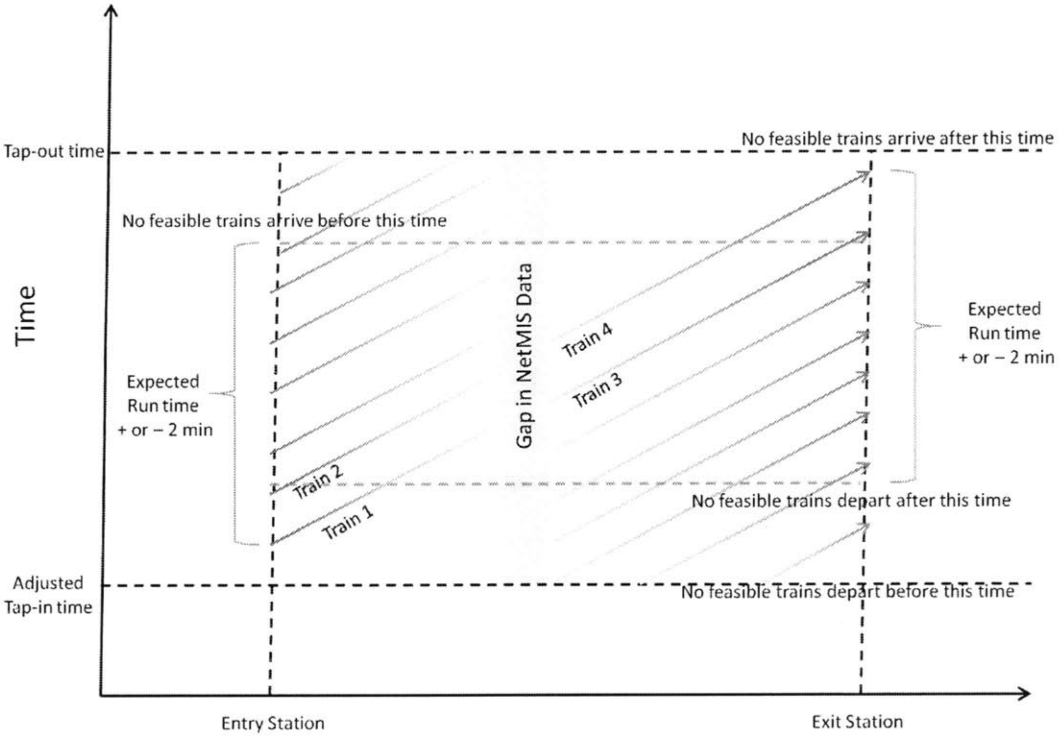

Time
Tap-out time
No feasible trains arrive after this time
No feasible trains arrive before this time
Expected
Run time
+ or – 2 min
Gap in NetMIS Data
Train 4
Train 3
Train 2
Train 1
Expected
Run time
+ or – 2 min
No feasible trains depart after this time
Adjusted
Tap-in time
No feasible trains depart before this time
Entry Station
Exit Station

Figure 3-13 Generation of Itineraries: Incomplete Train Identification Data with No Interchange

To generate path segments on lines that exhibit this behavior in the NetMIS data, and ultimately train itineraries on routes that involve these lines, the algorithm for generating itineraries is modified by adding a run time constraint and removing the matching Train Identification number constraint. In the case of a route that has no interchanges but is on an incomplete line, it is not necessary to generate itineraries because none of the path segments will be on a complete line. However, it is still useful to examine this case because it presents the modification to the algorithm in the simplest case.

Train Events for the entry and exit stations are collected as stated previously. The main modification is in the way Train Events at the entry and exit stations are linked. Previously, they were linked by train identification number and arrival and departure times, but because the train identification number is unreliable, in these cases the train identification number constraint is removed. Now any Train Event at the entry station can be linked to any Train Event at the exit station, as long as the entry station departure time is before the exit station arrival time. This creates a greater number of path segments, many of which will be infeasible because there is not enough run time between entry and exit station times. For this reason, an additional constraint is introduced to ensure that the time between the entry station and exit station is within two minutes of the expected run time for this leg. This expected run time comes from the RCM, one of the inputs to the Transit Network Process (section 3.5), and is readily available for each route. Adding this run time constraint prevents infeasibly short (or long) path segments being generated. The two minute window is selected arbitrarily.

In Figure 3-13, by adding the run time constraint, the only exit station trains that the entry station "Train 1" can be linked to are "Train 3" and "Train 4". This additional constraint results in a smaller feasible set of path segments.

Following is an example of generating itineraries for a passenger travelling between an OD pair with one possible route. This route has no interchanges and is on a line with incomplete train identification data. In table 3-12, an Oyster passenger enters at Acton Town at 8:32 (adjusted to 8:33) and exits at Earl's Court at 8:42. The expected run time from RCM for this OD pair is 7.5 minutes. Table 3-13 shows the three feasible itineraries generated from the transit network data for this passenger. There are only three pairs of Train Events that serve the route and satisfy the run time constraints within the Oyster passenger entry and exit times. Each feasible pair of Train Events is linked to form a train itinerary. Figure 3-14 shows the Oyster passenger's entry and exit times in relation to the three generated itineraries. It appears as though "Train" 1 and "Train" 3 or just "Train" 2 are feasible path segments because of their differences in run time.

PID	ENTRYSTN	EXITSTN	ENTRYTIME	EXITTIME
825674	Acton Town	Earl's Court	512	522

Table 3-12 Sample Oyster Record: Incomplete Identification Train Data with No Interchange

Train	Dep Station1	Dep Train Run1	Dep Train Time1	Arr Station1	Arr Train Run1	Train Time1	Run Time1
"Train" 1	Acton Town	17207030	513.92	Earl's Court	17231210	519.97	6.05
"Train" 2	Acton Town	17207030	513.92	Earl's Court	17230190	521.12	7.20
"Train" 3	Acton Town	17228460	514.67	Earl's Court	17230190	521.12	6.45

Table 3-13 Sample Generated Itineraries: Incomplete Train Identification Data with No Interchange

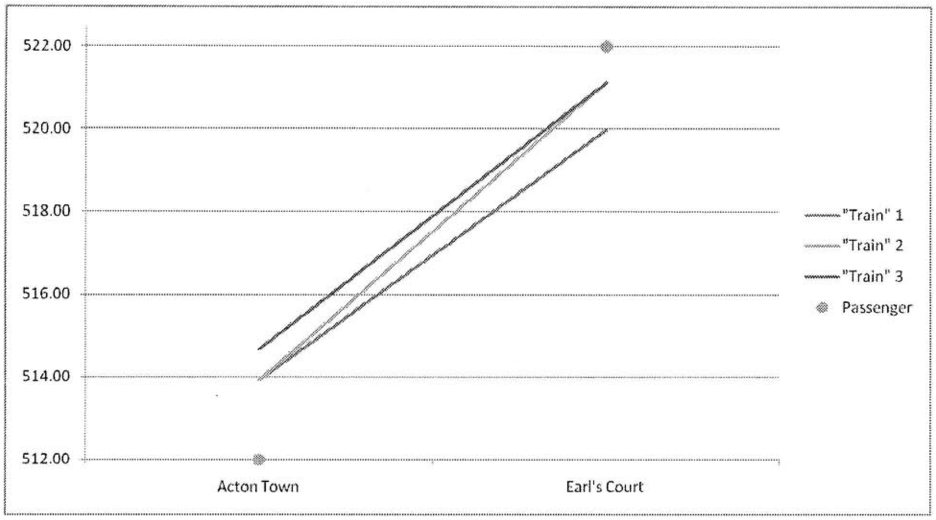


Figure 3-14 Sample Oyster Record with Generated Itineraries: Incomplete Train Identification Data with No Interchange

In summary, the algorithm used to generate itineraries for routes with incomplete train identification data and no interchanges is:

1. Extract all Train Events at the entry and exit stations between the passenger's entry and exit time.
2. Link Train Events at the entry station to those at the exit station by making sure that the entry station Train Event departure time plus the expected run time is within two minutes of the exit station Train Event arrival time.
3. Each pair of linked Train Events is a path segment and an itinerary.

Using the algorithm described above, figure 3-15 below shows the distribution of passengers by the number of feasible itineraries for passengers traveling on routes with incomplete train identification data and no interchanges. The highest percentage of passengers has one feasible itinerary, though many passengers have many itineraries. This spread distribution is a result of duplication of Train Events in path segments/itineraries, as shown in table 3-13.

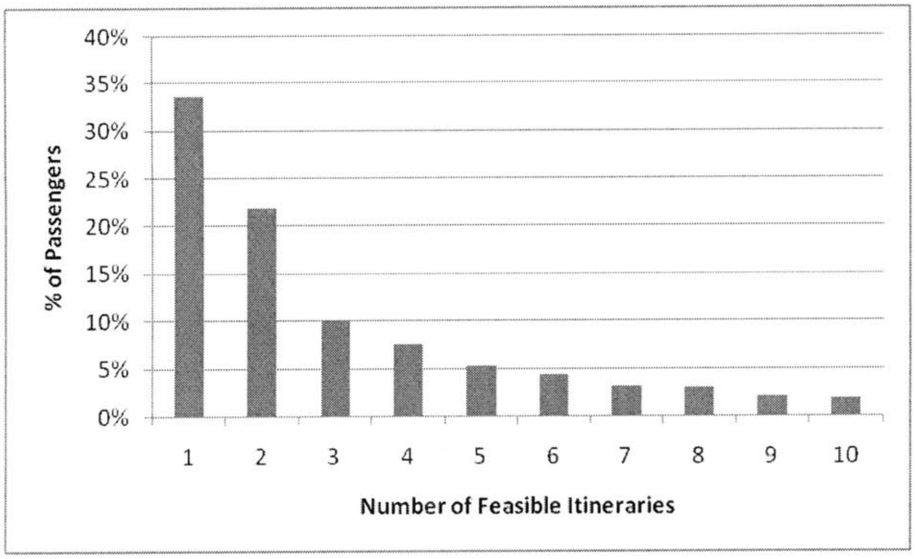


Figure 3-15 Distribution of Passengers by Number of Feasible Itineraries: Incomplete Train Identification Data with No Interchange

3.6.4 Generating Itineraries with Interchange and Incomplete Data

This section builds upon the concepts introduced in section 3.6.3 and introduces a new level of complexity: passenger trips involving interchange. This algorithm is crucial to the model's ability to compute comprehensive statistics on lines with complete data.

Essentially, for the legs of the route that are on lines with incomplete train identification data, path segments are generated by linking Train Events based on the relationships between arrival, departure, and run time. Then, infeasible path segments are eliminated as described in the algorithm for generating itineraries with interchanges by using path segments in adjacent legs. Finally, itineraries are generated by joining every feasible combination of path segments.

Figure 3-16 shows the process of generating itineraries for a route with one interchange and one line with incomplete train identification data. The first leg of the route is on a line with complete data and the second leg is on a line with incomplete data. The path segments for both legs are generated using the algorithms presented above. Infeasible path segments are then eliminated: the earliest possible arrival time at the interchange station constrains the feasible path segments on the second leg and the latest possible departure time from the interchange station constrains the feasible path segments on the first leg. Itineraries can then be generated based on the remaining path segments.

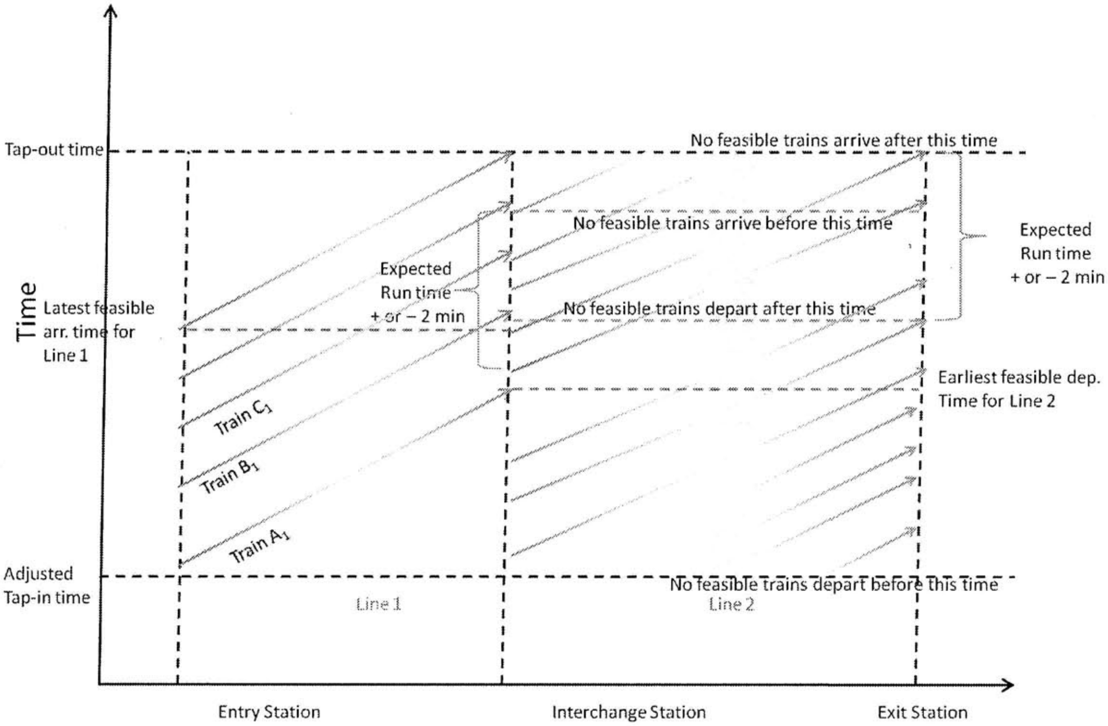

Time
Tap-out time
No feasible trains arrive after this time
No feasible trains arrive before this time
Expected Run time + or – 2 min
Expected Run time + or – 2 min
No feasible trains depart after this time
Latest feasible arr. time for Line 1
Earliest feasible dep. Time for Line 2
Train C1
Train B1
Train A1
Adjusted Tap-in time
No feasible trains depart before this time
Line 1
Line 2
Entry Station
Interchange Station
Exit Station

Figure 3-16 Generation of Itineraries: Interchange and Incomplete Train Identification Data

Following is an example of generating itineraries for a passenger travelling between an OD pair with a route that involves one interchange. This route has the first leg on a line with complete data and the second leg on a line with incomplete train identification data. In table 3-14, an Oyster passenger enters at Brixton at 9:01 (adjusted to 9:02) and exits at Waterloo at 9:20. In this example, only one route between Brixton and Waterloo will be considered: the Victoria line from Brixton to Stockwell for the first leg and the Northern line from Stockwell to Waterloo for the second leg. Table 3-15 shows all feasible path segments for the first leg, generated from transit network data between Brixton and Stockwell, between the passenger's entry and exit times. Table 3-16 shows all feasible path segments for the second leg, generated from transit network data between Stockwell and Waterloo, between the passenger's entry and exit times. Here, the train identification numbers do not match because the Northern line is a line with incomplete train identification data. Figure 3-17 shows the Oyster passenger's entry and exit times in relation to the four generated itineraries. The four generated itineraries are: Train 1 Leg 1 + Train 13 Leg 2, Train 1 Leg 1 + Train 14 Leg 2, Train 1 Leg 1 + Train 15 Leg 2, and Train 1 Leg 1 + Train 16 Leg 2. All infeasible path segments are eliminated.

PID	ENTRYSTN	EXITSTN	ENTRYTIME	EXITTIME
706081	Brixton	Waterloo	541	560

Table 3-14 Sample Oyster Record: Incomplete Train Identification Data with Interchange

Train	Dep Station 1	Dep Train Run1	Dep Train Time1	Arr Station 1	Arr Train Run1	Arr Train Time1	Run Time1
Train 1 Leg 1	Brixton	17181816	547.22	Stockwell	17181816	549.70	2.48
Train 2 Leg 1	Brixton	17206764	549.37	Stockwell	17206764	551.60	2.23
Train 3 Leg 1	Brixton	17209184	551.47	Stockwell	17209184	553.70	2.23
Train 4 Leg 1	Brixton	17207664	553.27	Stockwell	17207664	555.70	2.43
Train 5 Leg 1	Brixton	17229656	554.92	Stockwell	17229656	557.20	2.28

Table 3-15 Sample Path Segments for Leg 1: Incomplete Train Identification Data with Interchange

Train	Dep Station 2	Dep Train Run2	Dep Train Time2	Arr Station 2	Arr Train Run2	Arr Train Time2	Run Time2
"Train 1 " Leg 2	Stockwell	17233566	542.07	Waterloo	17229614	551.67	9.60
"Train 2 " Leg 2	Stockwell	17234302	542.98	Waterloo	17229614	551.67	8.68
"Train 3 " Leg 2	Stockwell	17234302	542.98	Waterloo	17238383	553.32	10.33
"Train 4 " Leg 2	Stockwell	17234302	542.98	Waterloo	17228234	553.62	10.63
"Train 5 " Leg 2	Stockwell	17224984	545.18	Waterloo	17238383	553.32	8.13
"Train 6 " Leg 2	Stockwell	17224984	545.18	Waterloo	17228234	553.62	8.43
"Train 7 " Leg 2	Stockwell	17234243	545.97	Waterloo	17238383	553.32	7.35
"Train 8 " Leg 2	Stockwell	17234243	545.97	Waterloo	17228234	553.62	7.65
"Train 9 " Leg 2	Stockwell	17234766	549.08	Waterloo	17222672	558.32	9.23
"Train 10 " Leg 2	Stockwell	17234766	549.08	Waterloo	17227516	558.67	9.58
"Train 11 " Leg 2	Stockwell	17225443	549.80	Waterloo	17222672	558.32	8.52
"Train 12 " Leg 2	Stockwell	17225443	549.80	Waterloo	17227516	558.67	8.87
"Train 13 " Leg 2	Stockwell	17235254	551.28	Waterloo	17222672	558.32	7.03
"Train 14 " Leg 2	Stockwell	17235254	551.28	Waterloo	17227516	558.67	7.38
"Train 15 " Leg 2	Stockwell	17236072	551.30	Waterloo	17222672	558.32	7.02
"Train 16 " Leg 2	Stockwell	17236072	551.30	Waterloo	17227516	558.67	7.37

Table 3-16 Sample Path Segments for Leg 2: Incomplete Train Identification Data with Interchange

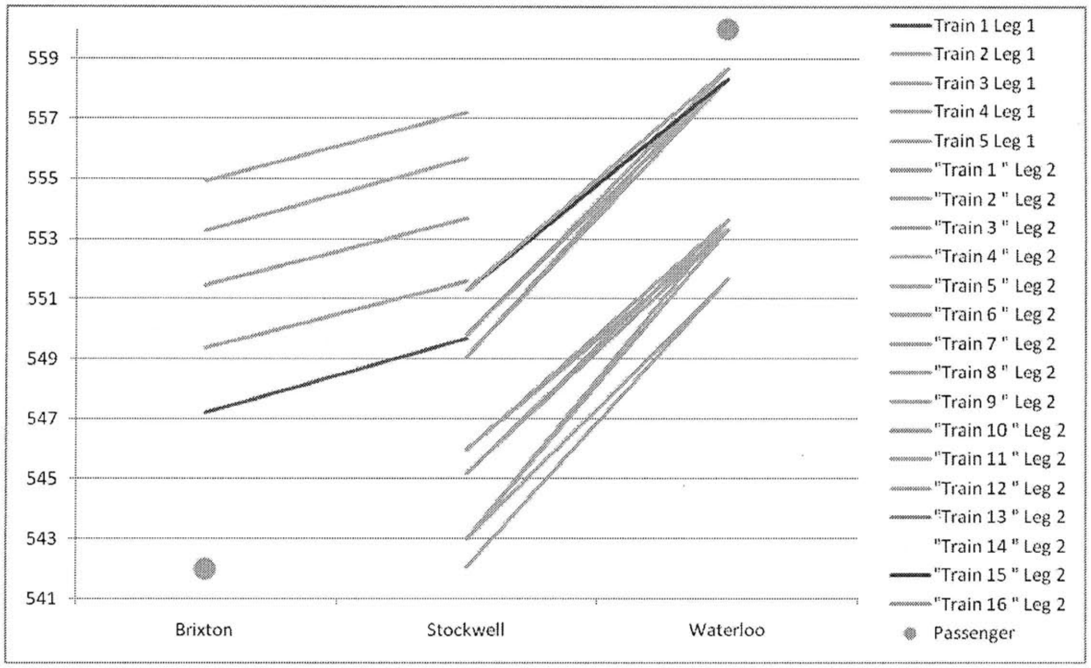


Figure 3-17 Sample Oyster Data with Generated Itineraries: Incomplete Train Identification Data with Interchange

In summary, the algorithm for generating itineraries for routes with incomplete train identification data and interchanges is as follows:

1. Retrieve all Train Events at the entry, exit, and interchange stations between the passenger's entry and exit time.
2. On legs with complete data, link Train Events making sure that the departure time from the start of the leg is before the arrival time at the end of the leg.
3. On legs with incomplete train identification data, link Train Events by making sure that at the start of the leg the departure time plus the expected run time is within two minutes of the arrival time at the end of the leg.
4. Each pair of linked Train Events is a path segment.
5. Eliminate all infeasible path segments using the earliest arrival times and latest departure times of path segments from adjacent legs on the route.
6. Generate itineraries by joining all feasible combinations of path segments.

Using the algorithm described above, figure 3-18 below shows the distribution of passengers by the number of itineraries for passengers traveling on routes with incomplete train identification data and interchanges. The highest percent of passengers still only have one itinerary, though many passengers have many itineraries. This spread distribution is a result of duplication of Train Events in path segments, and duplication of path segments in itineraries, as evident in table 3-16 and figure 3-17.

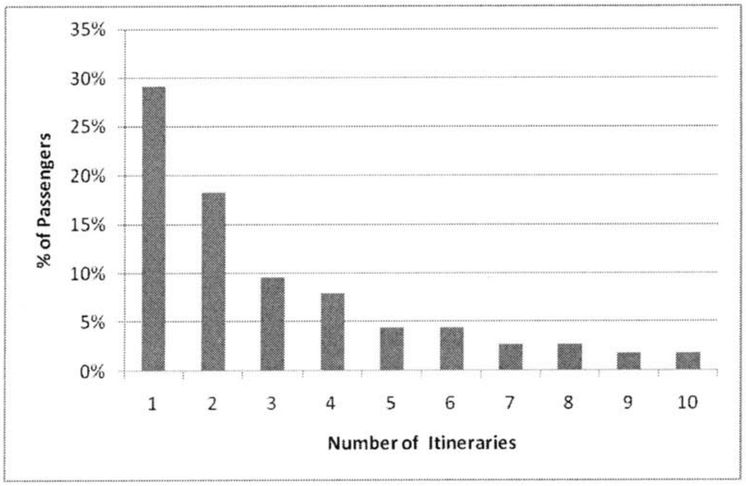


Figure 3-18 Distribution of Passengers by Number of Itineraries: Incomplete Train Identification Data with No Interchange

3.6.5 Incomplete Station Data

Sometimes, NetMIS fails to record any event logs for some stations in the LU network. All algorithms presented so far will fail if there are no Train Events at any end of any leg of a passenger's route since path segments cannot be created and therefore itineraries cannot be generated. However, it is important to generate itineraries for these routes because other legs might involve lines with complete data at some stations and the goal of this model is to produce results for these lines. The solution to this problem is to generate "dummy" path segments whenever there are no Train Events recorded for any leg of a route.

The underlying idea for the new algorithm presented in this section is illustrated in the following example: if the run time of the first leg is known, then the earliest possible time the passenger could have arrived at the interchange station, *t*, is known by adding the run time to the passenger's entry time. The passenger's journey can then be modified so that he appears to have started his trip at the interchange station with entry time as *t.* Given this, itineraries can be generated for this modified route. The same idea applies if a station with incomplete data is at the end of the route, or in the middle of the route. The run times for legs of routes, which come from the RCM, with incomplete station data are crucial in generating itineraries.

Instead of modifying a route each time such lines are encountered, "dummy" path segments are created. These dummy path segments do not contain actual train departure and arrival times, but supply the run time associated with that leg. These dummies are created so that each leg of a route can have path segments generated, regardless of the path segment being a dummy or not.

Similar to the previously presented algorithms, after generating the path segments for each leg of the route, the next step is to join the path segments together to generate itineraries. When a dummy path segment is joined to other path segments, it is assigned departure and arrival times. If this dummy path segment is at the start of a route, the departure time will be the passenger's entry time, and the arrival time will be the passenger's entry time plus the run time for that leg. If the dummy path segment is in the middle or the end of the route, the departure time will be the previous path segment's arrival time, and the arrival time will be the previous path segment's arrival time plus the leg's run time. In this way, path segments can be joined together chronologically and itineraries generated.

In summary, the algorithm for generating itineraries for routes with incomplete station data and interchanges is as follows:

1. Retrieve all Train Events at the entry, exit, and interchange stations between the passenger's entry and exit time.
2. Create path segments.
    a. On legs with complete data, link Train Events making sure that the departure time at the start of the leg is before the arrival time at the end of the leg.
    b. On legs with incomplete train identification data, link Train Events making sure departure time at the start of the leg plus the expected run time is within two minutes of the arrival time at the end of the leg.
    c. On legs with incomplete station data, create a dummy path segment that contains no data besides start and end station and run time.
3. Eliminate all infeasible path segments using the earliest arrival time or latest departure time of path segments from adjacent legs on the route. If the leg contains a dummy path segment, it will not be eliminated.
4. Generate itineraries by joining all feasible combinations of path segments.
    a. If two adjacent path segments are on lines with complete data or lines with incomplete train data, then they can be joined as long as they are in chronological order.

b. If one of the adjacent path segments is on a line with incomplete station data, then the dummy path segment either takes its departure time from the previous path segment, or from the passenger's entry time. It takes its arrival time from its departure time plus its run time.

3.6.6 Prioritizing and Combining Types of Incomplete Data

Unfortunately, with NetMIS data, incompleteness is seldom of just one type. Most lines besides the three complete lines (Victoria, Central, and Jubilee) have some combination of the two types of incompleteness. On lines with incomplete train data, many train identification numbers are not reassigned. On some lines with incomplete station data, event logs are generated some of the time at some stations. Given this, it is important to create path segments and generate itineraries that take full advantage of the availability and richness of data. Since it is impossible to predict the availability of data for every line, the algorithm must deal with all possible combinations of data incompleteness, and generate the best possible itineraries.

Path segments can be prioritized as follows:

1. Path segments with matching train identification numbers.
2. Dummy path segments and path segments with non-matching train identification numbers.

Dummy path segments and path segments with non-matching train identification numbers are given the same priority because neither is guaranteed to be more accurate than the other. Both depend on the expected run time, which may be different from the actual run time of the train.

For this reason, when a path segment for a leg on a line with incomplete (train or station) data is generated, the program first attempts to link non-matching train identification Train Events, if they exist. After the set of path segments for that leg are created, it is screened. First, the set is searched for path segments with matching train identification numbers. If such path segments exist, then the remaining path segments are deleted and the program moves onto the next leg. If not, it means the path segment set is either empty or contains only non-matching train identification number path segments. In either case, a dummy path segment is added to the set so that when the path segments for each leg are being matched chronologically, at least one of the path segments (dummy or non-matching train identification number) from the leg of interest will allow feasible itineraries to be generated. Finally, after all feasible itineraries are generated, if one (or more) of the itineraries contain no dummy path segments, the itineraries with dummy path segments are deleted. This process allows the program to produce a set of feasible itineraries for a route with incomplete data that is either comprised of all itineraries with non-dummy path segments, or itineraries that contain at least one dummy path segment.

In summary, the algorithm for generating itineraries for routes with all types of incomplete data is as follows:

1. Retrieve all Train Events at the entry, exit, and interchange stations between the passenger's entry and exit time.
2. Create path segments.
    a. On legs with complete data, link Train Events making sure that the departure time from the station at the start of the leg is before the arrival time from the station at the end of the leg.
    b. On legs with incomplete data, link Train Events making sure that the departure time at the start of the leg plus the expected run time is within two minutes of the arrival time at the end of the leg.
3. Screen the path segments on legs with incomplete data.
    a. If one (or more) path segment exists with matching train identification numbers, delete all other path segments.
    b. Otherwise, add a dummy path segment to the path segment set.
4. If the route contains an interchange, eliminate all infeasible path segments using arrival and departure times for adjacent path segments. If the leg contains a dummy path segment, it is not eliminated.
5. Generate itineraries by joining all feasible combinations of path segments ensuring that all path segments are in chronological order. If one of the adjacent path segments is a dummy path segment, then its departure time is taken either from the previous path segment, or from the passenger's entry time. Its arrival time is its departure time plus its run time.
6. Screen the list of feasible itineraries generated. If one or more itineraries exist with no dummy path segments, then delete all itineraries that contain dummy path segments. The set of feasible itineraries will contain either a list of itineraries with no dummy path segments, or a list of itineraries with at least one dummy path segment.

3.6.7 Oyster Clock Misalignment

This section deals with possible clock misalignment with Oyster data. The clocks at specific fare gates, which record entry and exit time, may be off by an unknown (but likely small) amount of time, and so incorrect times may be recorded, which in turn can lead to incorrect travel times. Incorrect travel times can lead to incorrect estimation of train itineraries for passengers, based on the proposed train itinerary generation methods. There is still one scenario where no feasible itineraries will be generated for all routes for a passenger's journey between a certain origin and destination: the time between entry and exit is shorter than the sum of all the run times for each leg of the route, for each route. The only explanation for this is misalignment of the clocks at the fare gates.

To correct for this situation, every time a passenger has no feasible itinerary after considering all possible routes, both the entry and exit times of the passenger trip are adjusted by one minute to extend the travel time, and the program again attempts to generate itineraries for that trip. This process is repeated up to five times if the program continues to find no feasible itineraries for the trip. The adjustment of plus or minus five minutes should be sufficient to correct for misalignment in the fare gate clocks.

3.6.8 Summary of Generation of Itineraries

The Generation of Itineraries process is summarized below:

1. Read each Oyster transaction from the Passenger Demand Process.
2. For each Oyster passenger, record entry station and time, and exit station and time.
3. Search for matching OD pair from output of the Transit Network Model.
    a. Return tree structure containing all possible routes, legs for each route, and all Train Events that serve each leg.
    b. If OD pair is not found, send error message and move on to next passenger.
4. Process each route associated with the OD pair.
5. For each route, read through each leg that makes up the route.
6. For each leg, determine if the line is complete or incomplete.
    a. If the line is complete, use the complete line algorithm to generate path segments.
    b. If the line is incomplete, use the incomplete line algorithm to generate path segments. This includes the screening process to remove inferior path segments.
7. Join path segments chronologically.

    a. If one of the path segments is a dummy, take the departure time from the previous path segment or the entry time, and the arrival time from the departure time plus the run time.

8. Remove all inferior itineraries (itineraries containing at least one dummy path segment) if possible.
9. Store final set of itineraries for each route.
10. Examine the number of feasible itineraries for each route.
    a. If none of the routes produce any feasible itineraries, use the Oyster Clock Misalignment mechanism.
11. Return the final set of feasible train itineraries for each route.

The key assumptions in Generation of Itineraries process are as follows:

1. A passenger's travel time on any leg of his journey is likely to be close to the expected run time for that leg.
2. The Oyster Clock Misalignment will be no greater than plus or minus five minutes and can be corrected for by lengthening the travel time by that amount.

3.7 Selection of Itinerary

After the generation of itineraries phase of the model, there will be more than one itinerary for many routes and OD pairs. This phase of the model takes as input the set of feasible train itineraries for each Oyster passenger transaction, and selects the most probable itinerary. There are two stages in selecting the most probably itinerary for a passenger. First, a passenger may have multiple routes, so a process must be developed to select the route that the passenger most likely took. Second, for any route, there may be multiple itineraries, so a process must be developed to select the most probable itinerary on that route. The process relies principally upon the input discussed in section 3.2.5: LU's average access, egress and interchange times. This section will describe the process to select an itinerary from the set of itineraries. Once an itinerary is selected for each passenger, statistics for the transit system can be calculated.

The goal in selecting an itinerary is to maximize the likelihood that this particular train itinerary and the particular route for the itinerary is the one the passenger actually travelled on. This decision is based on information revealed by the set of possible itineraries for each passenger. Several key pieces of information are known at the start of the selection process from the set of itineraries for each passenger. At the OD level, the following information is known:

1. **Number of Routes:** the number of routes that serve the passenger's OD and have feasible itineraries.
2. **Total Itineraries:** the total number of itineraries that are feasible for the passenger, regardless of route.

At the route level, the following information is known:

1. **Number of Itineraries:** the total number of feasible itineraries that serve the route.
2. **Classification of itineraries**: indication whether itineraries are based on complete data, or incomplete data, or include dummy path segments.
3. **Number of interchanges:** the total number of interchanges between the origin and destination.

At the itinerary level, the following information is known:

1. **Egress Time**: time between the arrival of the final train of the itinerary and the passenger's exit. In other words, it is the time a passenger took to exit the station given that he took this itinerary. This value is especially important because it is the only pure walking time value not containing waiting time. This value is used extensively in the itinerary selection process.
2. **Access and Platform Wait Time**: time between the passenger's entry time and the departure time of the first train of the itinerary. In other words, it is the time a passenger takes to pass through the gate, arrive at the platform, and board his train given that he took this itinerary. Access time includes platform wait time.
3. **Interchange and Platform Wait Time**: time between the arrival of the train at an interchange station and the departure of the next train of the itinerary, if an interchange exists. In other words, it is the time a passenger takes to alight his first train, walk to the interchange platform, and board his next train given that he selected this itinerary. Interchange time includes platform wait time.
4. **In Vehicle Time**: time between departure and arrival of the train on each leg of the itinerary.

This information is used to reduce the number of possible itineraries for a passenger, and select the most probable itinerary that a passenger travelled on. The rest of this section discusses the process of selecting the most probable itinerary in the following cases: single route with one feasible itinerary, single route with no interchange and multiple feasible itineraries, single route with interchange and multiple feasible itineraries, and multiple routes.

3.7.1 Single Route with One Itinerary: Creating Walk Time Distributions

In the case where a passenger has only one possible route with only one feasible itinerary, no selection is involved. These passengers, referred to as known itinerary passengers, comprise about 10% of all Oyster trips. These passengers are very much of interest, however, because their egress times can be measured with certainty—because the one feasible itinerary is certainly the itinerary that this passenger took, and the egress time associated with that itinerary is certainly the time the passenger took to exit the station. Thus, the egress times from these passengers can be used to create walk time distributions for each exit station. These walk time distributions will be used in the selection process for all other cases.

LU reports Average Values for Access, Egress and Interchange developed for the Journey Time Metric. The data for these average values were collected by surveyors walking all possible paths at 27 stations selected as a representative sample of all LU stations. All other LU stations are modeled using station and pedestrian flow simulation software. The walk times (a sample of five or so) collected at each station are averaged by time period. The average values for access, egress and interchange can be used to gain knowledge about the relationship between access, egress and interchange walk times at a station. These relationships will be used to help build distributions for access, egress and interchange times.

First, the egress times from known itinerary passengers who exited at the same station are grouped together, and distributions of egress time are created for each station, except for some stations in the network which have too small a sample size or no data at all. In this model, a minimum sample size of 75 is specified to be conservative. Because egress distributions need to be created at every LU station, stations with too a small sample size will be given an egress distribution from a similar station. LU categorizes stations into six types reflecting the passenger mix and volume: city, inner suburbs, outer suburbs, shopping, terminals and tourist stations. Table 3-17 shows the number of stations in each category, the number of stations sampled from passengers with known itineraries, the sample size of passengers with known itineraries for each category, the average egress time computed for each category, and the standard deviation of egress time computed for each category. Figure 3-19 shows the distribution of egress times from known itinerary passengers for all stations in each category. The probability density functions in figure 3-19 and the standard deviations from table 3-17 show that there is a great deal of variation among stations within each category. Some of this variation may be due to the Oyster timestamp truncation.

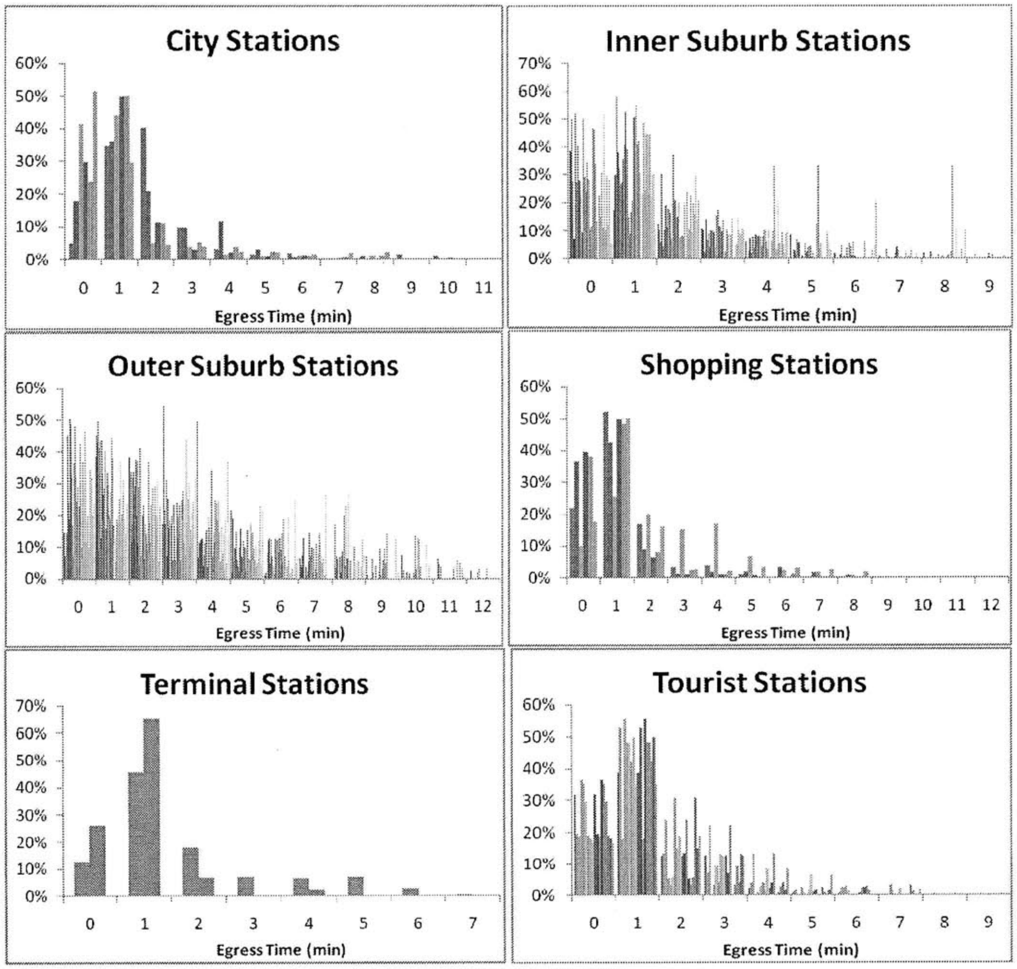
Grouping stations by station category may not be the most intuitive way to find stations with similar egress distributions. An alternative might be to group stations by structural characteristics: deep tube stations, sub-surface stations, number of stairs and escalators, and layout of platforms. Because of time pressures, alternative methods for grouping stations were not explored.

Category	Count of Total Stations	Count of Sampled Stations	Known Itinerary Passenger Sample Size	Average Egress Time	Standard Deviation of Egress Time
City	20	6	6163	1.45	1.58
Inner Suburb	76	40	4587	1.77	1.81
Outer Suburb	135	86	4497	2.50	2.36
Shopping	9	6	3970	1.30	1.49
Terminus	7	2	568	1.82	1.55
Tourist	21	9	2004	1.35	1.47

Table 3-17 Station Category Statistics

It is important to note that the expected value of the egress distribution at a particular station is not necessarily the same as LU's average egress time. Figure 3-20 plots the average egress times from known itinerary passengers against LU's standard egress values. LU's values should be slightly larger than the measured egress time because LU measures egress as the time from the platform to the exit of the station while this model measures egress as time from the platform to the fare gate. For this reason, most points in this graph should be above the 1:1 ratio line. The blue dots are those stations that LU had surveyed manually. Nearly all of these stations are above the 1:1 ratio line. The fact that nearly all stations with manual surveys appear to be in line with the estimated egress time indicates that this model accurately estimates egress time. The remaining stations are represented with red dots. The fact that many stations with egress times that were not manually surveyed by LU are below the 1:1 ratio line suggests that LU's estimations of egress times may be underestimated. For all stations manually sampled by LU, the average of the LU standard egress times is 186 seconds. For those same stations, the average of this model's average egress times 121 seconds. The ratio of LU's average to this model model's average is 1.53. For all stations not sampled by LU, the average of LU standard egress time is

108 seconds. For those same stations, the average of this model's observed egress time 186 seconds. The ratio LU's average to this model model's average is 0.58. The difference in these two ratios exemplifies the potential difference in accuracy between the LU sampled stations, and all other stations.



3-19 Distributions of Egress Time for Known Itinerary Passengers

With observed egress distributions for every station, the relationship between LU's average access and egress times can be used to build access distributions for every station in the LU network. The key assumption in building an access time distribution is that the ratio between LU's average egress time and access time will be the same as the ratio between the expected value of the egress and access distributions. Once the ratio between LU's average egress and access time is calculated, the expected value of the egress distribution can be used to solve for the expected value of the access distribution. Taking this idea one step further, if the expected value of the access distribution at a particular station could be determined by applying the LU ratios to the expected value of the egress distribution at a particular station, then every value in the access distribution could be determined by applying the LU ratios to every value in the egress distribution. This process effectively scales the egress distribution at a particular station by the LU ratios and creates an access distribution and is described mathematically below.

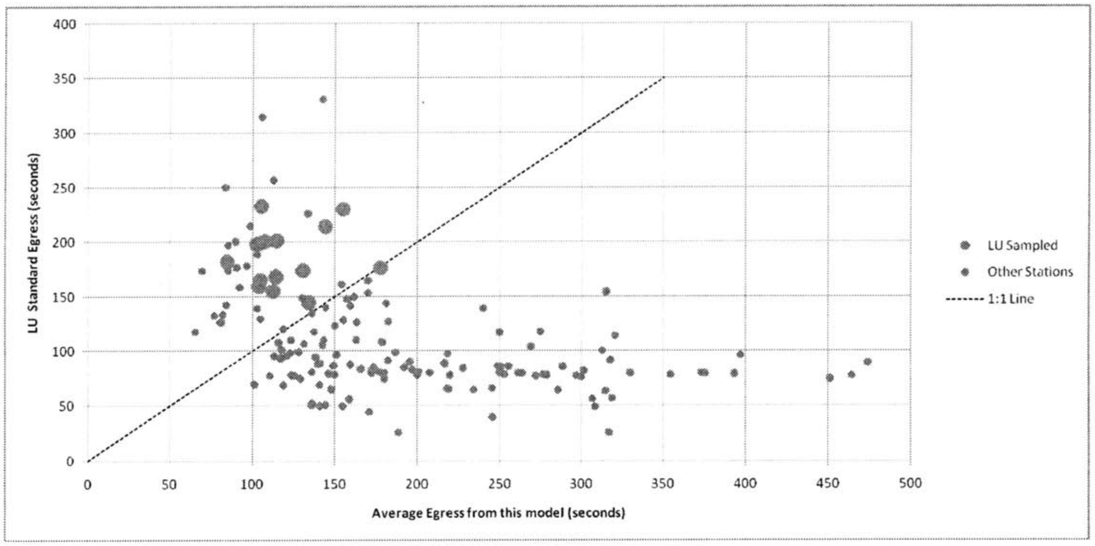
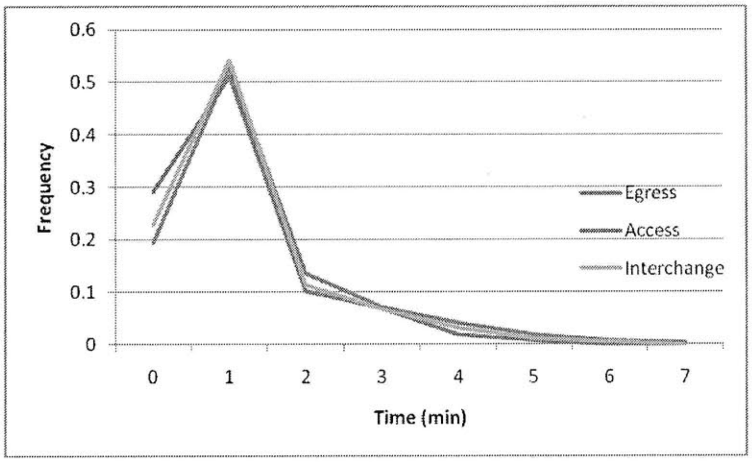


Figure 3-20 Average Egress from Known Itinerary passengers vs. LU Standard Egress Values

To create an access time distribution at station i:

- Let A_i represent LU's average access value at station i.
- Let E_i represent LU's average egress value at station i.
- Let $r_i^{AE} = \frac{A_i}{E_i}$, the ratio of LU's average access and egress values at station i.
- Let N_i represent the number of egress time observations in the egress distribution at station i.
- For each egress time observation, $e_{i,n}$, at station i, calculate the corresponding access time observation, $a_{i,n}$, as follows: $a_{i,n} = e_{i,n} * r_i^{AE} \quad \forall n \in N_i, \ i = 1,2, \dots k$
- Therefore, access time is computed by multiplying the ratio of LU's average access and egress times by the observed egress time.

Interchange time distributions are created in an analogous manner. Figure 3-21 shows the egress distribution from known itinerary passengers for Green Park station, and the inferred access and interchange time distributions. The ratio of LU's average egress time to average access time at Green Park is 0.98 and the ratio of average egress to average interchange at Green Park is 1.16. When grouped in one-minute bins, the distributions appear to be very similar.



3-21 Access, Egress and Interchange Distributions for Green Park

This program is run once to collect the egress time observations from known itinerary passengers. Then, access and interchange distributions are created through the process described above. These distributions for access, egress and interchange time for each station are then stored in a hash table structure with a key for every station on the transit network and values that contain the three distributions. These distributions are then used as input to future iterations of the model, and are used to select the most probable itinerary for passengers in the other cases.

Another method considered for generating access, egress and interchange distributions involved building egress distributions as described above, and creating Access + Platform Wait Time distributions for each station and isolating the access time by subtracting the expected waiting time at each station (based on train headway). However, this method was found to be complex and difficult to automate, failing in some cases where the expected wait time was greater than the access + platform wait time.

These access, egress and interchange distributions may be more accurate if they are reported to a finer level of detail, for instance, by line and direction. For example, at an interchange station with a deep tube line and a sub-surface line, the access distributions may be very different by line. Passengers using the deep tube line would have longer average access times compared with passengers using the sub-surface line. There appears to be sufficient data for this kind of analysis to be successful. This idea has not been implemented because of time pressures.

3.7.2 Single Route with No Interchange and Multiple Feasible Itineraries

To select an itinerary for passengers with one possible route with no interchanges and multiple feasible itineraries, the generated access and egress time distributions are used to reduce the set of itineraries.

The approach to reducing the number of possible itineraries is based on a hypothesized relationship between a passenger's egress and access time. Assuming that a passenger walks at the same speed at his entry station and exit station leads to the notion that itineraries with similar walking times at each station should be chosen. The access and egress time distributions provide a sufficient basis for comparison of walking times for every itinerary for specific origin and destination stations. Similar access and egress times can be described as times that correspond to the same percentiles in their respective distributions. In order to compare alternative train itineraries, a method is developed to determine the itinerary with the most similar access and egress times.

This similar walk time assumption can be described as follows: a passenger's access time should be in the same percentile of the cumulative access time distribution at the entry station as the egress time is in the distribution at the exit station. Accordingly, the program evaluates each itinerary as follows:

1. Find the percentile of the estimated egress time in the cumulative egress time distribution at the exit station.
2. Obtain the access time associated with the same percentile in the access time distribution at the entry station.
3. If there is enough time between the entry time and the departure time of the first train of the itinerary, then the itinerary passes this filter. In other words, if the access plus platform wait time is greater than or equal to the inferred access time, the itinerary is deemed to be feasible.
4. The difference between the two values is defined as the platform wait time at the entry station.
5. Itineraries which do not allow the inferred access time are eliminated.

There are two circumstances where this process will fail to reduce the set of feasible itineraries to one:

1. The process is too restrictive: when passengers' access time was much faster relative to their egress time for each itinerary, so all feasible itineraries are eliminated. In this case, the program will revert to the original set of feasible itineraries.
2. The process is not restrictive enough: when more than one itinerary has sufficient access time built into the itinerary and more than one itinerary passes the screen. In this case, the surviving set of feasible itineraries is retained.

In both circumstances, one final step is needed to reduce the set of possible itineraries to one: to select the train itinerary with the most probable egress time. This is done by finding the probability of each itinerary's egress time from the egress time distribution. If two (or more) train itineraries have the most probable egress time, then an itinerary is randomly selected from that set. At this point, there is only one itinerary left and the selection process is complete.

Table 3-18 shows an example of a set of itineraries for the passenger shown in table 3-7 having one possible route with no interchange. The first itinerary's egress time falls at the 93rd percentile of the pre-loaded egress distribution at Victoria. The second itinerary's egress time falls at the 21st percentile. The 93rd percentile of the access distribution at Brixton is 5.21 minutes, and the 21st percentile is 0.75 minutes. The first itinerary does not allow enough time in the Access + Platform Wait Time for 5.21 minutes of access time. The second itinerary does allow enough time for 0.75 minutes of access time. The remaining time, 4.65 minutes, is designated as platform wait time. The first itinerary is eliminated and the second itinerary is selected. The process for determining whether this passenger is classified as being left behind will be described in section 3.8.

Itinerary	Dep Station1	Dep Train Run1	Dep Train Time1	Arr Station1	Arr Train Run1	Arr Train Time1	Run Time1	Access Time + PWT	Egress Time
1	Brixton	17202313	527.57	Victoria	17202313	536.77	9.20	3.57	2.23
2	Brixton	17204291	529.40	Victoria	17204291	538.72	9.32	5.40	0.28

Table 3-18 Set of Itineraries for Route with No Interchange

Other methods for reducing the number of feasible itineraries by applying access and egress times involved requiring that each passenger's access and egress times were equivalent to LU's average access and egress times and eliminating itineraries that were deemed infeasible on this basis. This rule and similar variations of this rule proved to be too blunt and eliminated too many feasible itineraries. The method presented above is a more sensitive approach because it takes into account a passenger's walking patterns.

Still, this method may not be sensitive enough. Translating the assumption that passengers have similar walking times at their entry and exit stations into having matching percentiles in their access and egress stations is a bit rigid. A passenger might hurry through a station at the start of his journey (in order to catch an earlier train) and slow down at his exit station because he is not as late as he thought he might be. Future work should include development of a method of applying access and egress times that is less rigid than the method presented above.

3.7.3 Single Route with Interchange and Multiple Feasible Itineraries

To select an itinerary for passengers with a single route with interchange and multiple feasible itineraries, the generated access, interchange and egress time distributions are also used to reduce the set of feasible itineraries. The process for these passengers is very similar to the process presented in section 3.7.2 for non-interchange passengers:

1. Egress and access time distributions are used to reduce the set of itineraries.
2. If the set is reduced to a single itinerary, no further action is necessary.
3. If all feasible itineraries are eliminated, the program reverts to the original set of itineraries.
4. If there is more than one remaining itinerary, the program repeats steps 1 and 2, replacing the original set of itineraries with the reduced set, and the access distribution with the interchange distribution.
5. If all itineraries are eliminated, the program will revert to the previous set of itineraries.
6. If there is more than one itinerary and more than one interchange, the program will repeat steps 4 and 5 for each interchange station sequentially.
7. If all interchange stations are exhausted and there is still more than one itinerary, the itinerary with the most probable egress time is selected.

Table 3-19 shows an example of a set of itineraries for the passenger in table 3-9 having a single route with one interchange. The first and third itineraries' egress times fall at the 68th percentile of the egress time distribution at Leicester Square. The second itinerary's egress time falls at the 88th percentile. The value at the 68th percentile of the access time distribution at Victoria is 0.78 minutes, and the 88th percentile is 1.67 minutes. The second itinerary does not allow enough time in the Access + Platform

Wait Time for 1.67 minutes of access time, and is therefore eliminated. The first itinerary allows enough time for 0.78 minutes of access time with 0.72 minutes of platform wait time. The third itinerary allows enough time for 0.78 minutes of access time with 2.77 minutes of platform wait time.

Interchange times are then applied to the reduced set that contains the first and third itineraries. The value at the 68th percentile of the interchange distribution at Green Park is 1.42 minutes. The interchange time in the third itinerary does not allow enough time for 1.42 minutes, and it is therefore eliminated. The first itinerary allows enough time for 1.42 minutes of interchange time, with 1.43 minutes of platform wait time. The first itinerary is the only remaining itinerary and is selected.

Itinerary	Dep Station 1	Dep Train Run1	Dep Train Time1	Arr Station 1	Arr Train Run1	Arr Train Time1	Run Time 1	Access Time + PWT
1	Victoria	17237824	578.50	Green Park	17237824	580.63	2.13	1.50
2	Victoria	17237824	578.50	Green Park	17237824	580.63	2.13	1.50
3	Victoria	17224454	580.55	Green Park	17224454	582.63	2.08	3.55

Itinerary	Dep Station 2	Dep Train Run2	Dep Train Time2	Arr Station 2	Arr Train Run2	Arr Train Time2	Run Time 2	Interchange Time + PWT	Egress Time
1	Green Park	17229820	583.48	Leicester Square	17229820	587.13	3.65	2.85	0.87
2	Green Park	17229550	581.95	Leicester Square	17229550	585.13	3.18	1.32	2.87
3	Green Park	17229820	583.48	Leicester Square	17229820	587.13	3.65	0.85	0.87

Table 3-19 Set of Itineraries for Route with Interchange

3.7.4 Multiple Routes

All previous cases have dealt with passengers travelling on OD pairs served by only a single route. In the case where a passenger has multiple routes, there might be a set of possible itineraries for each route. In order to increase the efficiency of the program, the program selects the preferred route prior to examining each itinerary to the level described in the previous cases. The route is chosen by examining the number and type of possible itineraries per route.

Figure 3-22 shows how a passenger might have different numbers of feasible itineraries for each route that serves the OD pair in question. By choosing the route prior to examining each itinerary, there can be major savings in computation time because the number of itineraries to be examined is reduced with each route that is eliminated.

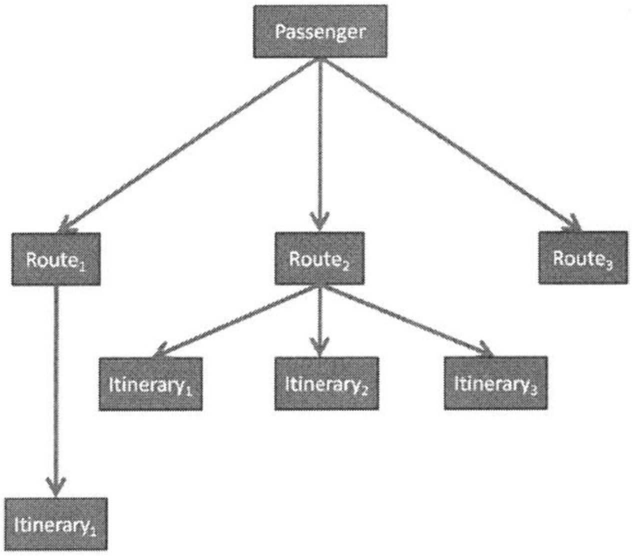


Figure 3-22 Passenger with Multiple Routes

The first premise in choosing a route is that this choice is made before the passenger arrives at the platform of a station. This is generally true in the LU transit network because most lines do not share platforms with other lines, so the passenger must choose his route in order to decide at which platform to wait for a train. Once a passenger has arrived at a platform, his choice set consists only of train itineraries belonging to the chosen route. For this reason, it makes sense to compare the passenger's choice set at the route level before comparing the individual qualities of each itinerary against all other itineraries.

Routes can be compared by considering the number and type of itineraries for each route. Because train itineraries are generated based on the passenger's travel time, the number of itineraries for each route may be related to how well the route fits the travel time. If a route requires a run time longer than the passenger's travel time, no feasible itineraries will be generated for that route. Conversely, if a route requires a run time that is much smaller the passenger's travel time, many itineraries will be generated for that route. For a selected route, the number of itineraries generated for that route should be very small because the passenger's travel time should closely fit the run time of the route. Based on this, one criterion for choosing the route is that the best route should have the smallest number of itineraries greater than zero.

However, the number of itineraries for a route will not always indicate whether or not a passenger took that route. For example, when a passenger chooses a short but crowded route and is left behind by several trains (each representing a different itinerary), the set of feasible itineraries for his chosen route will be large. If this passenger has a less crowded but longer route available, his set of possible itineraries for this route may be small. Under the stated selection criterion, the route selected will be the longer but less crowded route. Therefore, a corollary of this premise is that a passenger will choose the route in which he is less likely to be left behind: effectively the less crowded route. The implication of this assumption can be tested given the results of this model and the estimated level of crowding on trains for each route can be examined in future work.

One final criterion for route selection is the type of train itinerary. A set of itineraries for a route either consists of all train itineraries having dummy path segments, or no dummy path segment itineraries. In order to maximize the use of the available NetMIS data, routes with no train itineraries containing dummy path segments are chosen over routes with train itineraries containing dummy path segments, regardless of the number of possible itineraries.

Once the route is chosen, the passengers fall under one of the cases previously presented and the most probable itinerary can be selected by applying access, egress and interchange times as discussed above.

3.7.5 Summary of the Itinerary Selection Process

The Itinerary Selection Process selects a train itinerary by applying the following rules:

1. Read in the set of possible train itineraries produced by the Generation of Itineraries process. This set of itineraries corresponds to all the possible combinations of trains a passenger could take to travel between a certain OD pair during a certain time, for each route between the OD pair.
2. If the set contains only one itinerary, that itinerary is selected and this process ends.
3. Select the best route based on the following rules:
    a. A route with path segments based on complete service data and the fewest itineraries.
    b. A route containing path segments based on incomplete service data and the fewest itineraries.
4. If there is only one itinerary for the selected route, that itinerary is selected and this process ends.

5. Reduce the number of itineraries by applying access and egress times at the entry and exit stations.
    a. If all itineraries are eliminated, revert to the original set of itineraries and move onto the next step.
    b. If there is only one itinerary left, the itinerary is selected and this process ends.
6. For the remaining itineraries containing at least one interchange, reduce the number of itineraries by applying interchange times and egress at the first interchange and exit stations.
    a. If all itineraries are eliminated, revert to the previous set of itineraries and move onto the next step.
    b. If there is only one itinerary left, the itinerary is selected and this process ends.
7. For the remaining itineraries containing at least two interchanges, reduce the number of itineraries by applying interchange times and egress at the second interchange and exit stations.
    a. If all itineraries are eliminated, revert to the previous set of itineraries and continue this process up to four interchanges.
    b. If there is only one itinerary left, the itinerary is selected and this process ends.
8. If there is still more than one itinerary left, select the itinerary with the most probable egress time.

The key assumptions in the itinerary selection process are:

1. A passenger chooses his route prior to arriving at the platform.
2. A passenger will choose the longer, less crowded route over a route where he is likely to get left behind by a train due to crowding.
3. A passenger walks at the same speed at his entry, exit, and interchange stations.
4. The ratio between LU's average access, egress and interchange times are proportional to the ratio between the expected value for the distributions for access, egress and interchange times for each station.
5. Stations in similar categories (tourist, city, etc.) have similar distributions for access, egress and interchange time.
6. All else equal, the itinerary with the most likely egress time is the itinerary the passenger selected.

3.8 Transit System Statistics

This final process in the model takes the selected passenger itinerary data and reports the load on each train at each station on the Victoria and Jubilee lines; the average access, egress and interchange times at each station; the number of passengers left behind at each station; and the relationship between load and the probability of being left behind. The Central line is excluded from this report because its loop and multiple branches caused the computation of load be more difficult than other lines and infeasible because of time pressures.

Updating Load and Access, Egress and Interchange Times

First, every time an itinerary is selected, all passengers assigned to that itinerary are added to the load of each train of that itinerary. When the same train is selected by another group of passengers, the load is updated.

Then, the egress time from that itinerary is added to the list of egress observations at the exit station. This automatically updates the list of access and interchange observations since they are based on the egress distribution. The process up updating AEI times is done to provide richer distributions for each station and moves the program away from dependency on the similar station criteria when there are not enough egress observations for a station. There is a risk involved with this step: if this process is incorrectly selecting train itineraries for passengers, then the incorrect egress times are being updated to the egress (and access and interchange) distributions. Future research might investigate different model results if the AEI distributions were not updated.

Counting the Number of Trains That Leave a Passenger Behind

A left behind passenger is defined as a passenger that did not board the first train that arrived at the entry station after he arrived at the platform. There are three possible explanations for this:

1. The passenger was unable to board the train because of crowding.
2. The passenger chose not to board the train.
3. The passenger actually arrived at the platform after the train departed (the passenger's access time was incorrectly estimated).

To determine if a passenger was left behind by a train, the inferred egress time and the corresponding inferred access and interchange times from the selected itinerary are used to determine the passenger's arrival time at each station of the journey. First, all itineraries are checked to make sure that they allow enough walk time, as defined by the selected itinerary. This is done by applying the walk times to the set of possible itineraries. It is important to note that the set of possible itineraries are all from the same route. If these itineraries are feasible given the walk times of the selected itinerary, then these candidate left behind itineraries must be examined path segment by path segment to check if trains in the candidate itineraries departed after the passenger's arrival times at the relevant stations, but before the departure time of the trains from the selected itineraries. If trains from candidate left behind itineraries fit this criterion, then they are counted at each station of the passenger's journey.

For example, the selected itinerary for the passenger in table 3-7 is the second itinerary in table 3-18. The inferred access time for the passenger is 0.75 minutes. This means the passenger is inferred to have arrived at the platform at 526.75 minutes past midnight (8:46:45 AM), which is before the departure time of the train from the first itinerary. The passenger is determined to be left behind by the train from the first itinerary.

Compiling the Data

Finally, a table is written that contains relevant data from every Oyster transaction, data from the selected itinerary including train identification numbers, arrival and departure times and station, egress time, and the number of trains by which the passenger or passengers were left behind at each station of the journey. After all passengers are assigned to an itinerary and therefore a set of trains, the program iterates through each passenger again, and reports the load on the train(s) the passenger boarded at the time of boarding.

4 Model Application

This chapter presents the results from the application of this model to the LU network. The model infers which train itinerary a passenger took to get from his origin to destination through the processes described in chapter 3. After a train itinerary is selected for each passenger, various figures and statistics are calculated to assess the performance of the network, as well as the credibility of the model. The model takes input from the London Underground network and therefore the model's accuracy is evaluated relative to other knowledge about this network when possible. The first section of this chapter discusses how the model is presented so that the accuracy of the model can be evaluated. The following sections present the various results of the model, according to the structure outlined in the first section. These results include the number of possible itineraries per passenger, the percent of passengers left behind, and this model's predicted train loads. The final section presents a brief sensitivity analysis of the results.

4.1 Structure of Results

Results produced from this model are presented so as to help assess its accuracy. In terms of passenger-centric results, there are many elements in a passenger's journey that need to be tested for accuracy in order to understand the model's strengths and weaknesses. For example, it is important to examine how precisely the model identifies the train a passenger chose when there are no interchanges in his journey, compared to when there are interchanges. Similarly, the percent of passengers "left behind" by crowded trains must be compared against the percent of passengers being left behind by uncrowded trains. For this reason, passengers whose journeys are of certain types are grouped into four subsets and results are presented separately for each subset of passengers, as explained below. The remaining results that pertain to trains and the transit network will be presented in a different manner. The results will be presented only for the lines with complete NetMIS data: Victoria and Jubilee lines. Though the Central line has complete data, it is excluded because efforts to compute load on lines with branches and loops proved difficult and not feasible within the time constraints of this research.

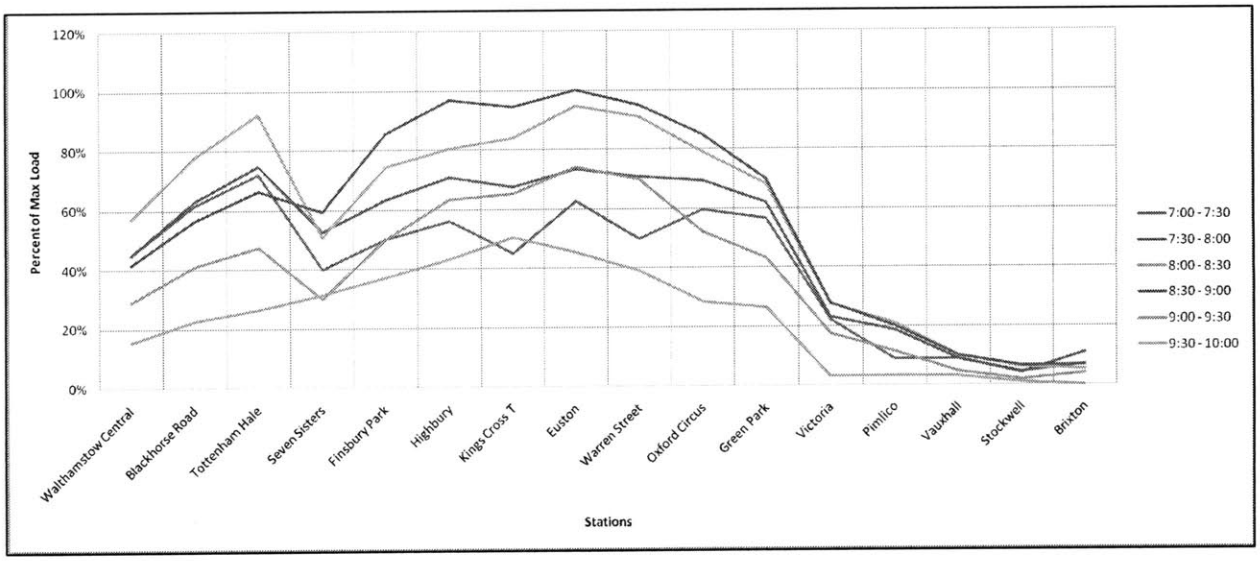
4.1.1 Subset 1: Control Group

This subset contains passengers that travel between OD pairs that have:

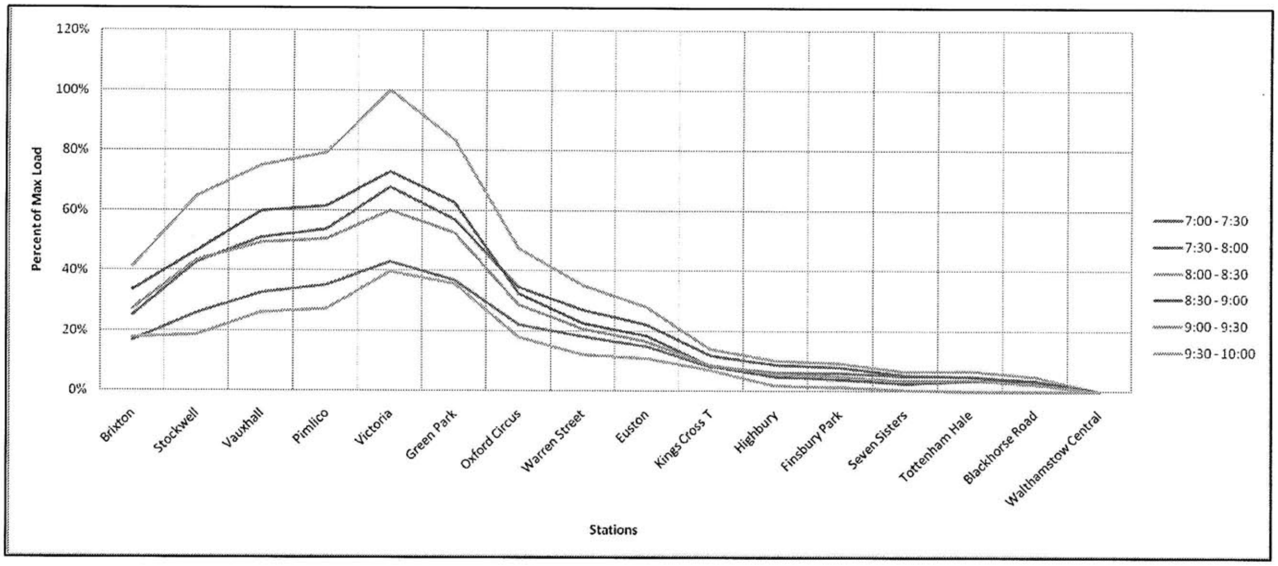
1. One likely route
2. No interchange

3. Low level of congestion when boarding

The number of possible routes and the existence of interchanges can be determined through the Transit Network Model for each OD pair. The first two constraints eliminate all OD pairs except those served by only one line in the transit network, which are OD pairs whose best paths include no interchanges. Only trips on the Victoria and Jubilee lines are considered. The constraint regarding the level of congestion can be assessed by examining the load on the train that the passenger boarded. A low level of congestion is defined as less than 25% of the maximum observed load. The maximum observed load is the highest load on any train on the same line and direction as the train in question. Figures 4-1 and 4-2 show examples of the load profile on the Victoria line. These figures are produce by the model. The maximum observed loads have similar values in both directions.



4-1 Load Profile for Victoria Line Southbound



4-2 Load Profile for Victoria Line Northbound

Figure 4-1 shows the load on the Victoria line southbound trains. The load is calculated by adding passengers to each train in their selected itinerary. Each line presents the average load of trains that depart during the given 30-minute period. For each time period, the load at each station is represented as a percent of the maximum average load of southbound trains on the Victoria line across the entire AM peak period. In this example, if each line represented a single train, the maximum southbound load on the Victoria line would be during the 8:30-9:00 AM time period at Euston station. Figure 4-2 similarly represents the average load at each station for each time period on the Victoria line northbound. The maximum northbound load would be during the 8:00-8:30 time period at Victoria station. Passengers that start at Victoria, or any station further south, and travel south (with the exception of passengers that start their trips between 8:30 and 9:00) are included in subset 1. Passengers that start between 8:30 and 9:00 must start at Pimlico or any station further south and travel south in order to be included in subset 1.

By including in this subset only passengers that travel under these clearly uncongested conditions in this subset, it becomes easy to isolate and study the model's ability to assign passengers to their correct trains without any interference from issues such as route choice, interchange, and congestion. It also establishes the base expectation for the performance of the model. It represents the simplest passenger trips, so the model should provide the most accurate results for this group. Any error in results observed in this subset will only be compounded for the other subsets of passengers. For this reason, subset 1 can be seen as the control group.

Passengers in this subset make up approximately 33% of all passengers that travel exclusively on the Victoria and Jubilee lines.

4.1.2 Subset 2

This subset contains passengers that travel between OD pairs that have:

1. One likely route
2. No interchange
3. High levels of congestion upon boarding

These passengers are similar to subset 1 except that it includes only passengers that may experience high levels of congestion. For subset 2, a high level of congestion is defined as any load greater than 80% of the observed maximum load.

By limiting this subset to passengers under these constraints, it becomes easy to isolate and study the model's ability to assign passengers to their correct trains given interference from congestion. It also helps to identify how the model handles passengers who face congestion when comparing the results from subsets 1 and subset 2.

Passengers in this subset make up approximately 3% of all passengers that travel exclusively on the Victoria and Jubilee lines.

4.1.3 Subset 3

This subset contains passengers that travel between OD pairs that have:

1. One likely route
2. Interchange
3. Low levels of congestion upon boarding

These passengers are similar to subset 1 except that it only includes passengers that travel on OD pairs that require an interchange, typically meaning that the origin and destination are on different lines. Passengers must not face congestion on any leg of their journey. Only trips that involve interchange between lines with complete NetMIS data will be considered. By considering these passengers in this subset, it becomes easy to isolate and study the model's ability to assign passengers to their correct trains when interchange is involved.

Passengers in this subset make up less than 1% of all passengers that travel exclusively on the Victoria and Jubilee lines.

4.1.4 Subset 4

This subset contains passengers that travel between OD pairs that have:

1. More than one route
2. Possible interchange
3. Low levels of congestion upon boarding

The difference between subset 4 and subset 1 is that subset 4 only includes passengers that travel on OD pairs that have multiple possible routes. Passengers must face low load on every leg of their journeys. There is no restriction regarding the existence of interchange. Only trips that involve lines with complete NetMIS data will be considered.

By including passengers under these constraints in this subset, it becomes easy to isolate and study the model's ability to assign passengers to their correct trains when there are multiple possible routes. There is no restriction in terms of interchange because there are few OD pairs that have a path choice set with no interchange. Often, the path choice set has some paths with interchange, and others without. This subset directly tests the route choice assumption in the model discussed in chapter 3.

Passengers in this subset make up approximately 1% of all passengers that travel exclusively on the Victoria and Jubilee lines.

The remaining passenger trips not included in these four subsets possess some combination of the "extreme" qualities described for each subset, or have no singular distinguishing quality (high load, low load, etc). These trips will not be grouped into a subset and studied because they do not possess a single quality that would be interesting to study, nor are there any a priori expectations of the results for these passenger trips. For example, for passengers experiencing medium loads on their trips, there is no reason to expect that they should have a high or low rate of being left behind, or a high or low number of feasible itineraries. Another example is passengers whose trips involve interchange, and experience mixed levels of congestion. It would be impossible to distinguish the effects of interchange and congestion on their results. For these reasons, no separate results will be presented on the passengers not included in subsets 1, 2, 3 or 4.

4.1.5 Structure of Other Results

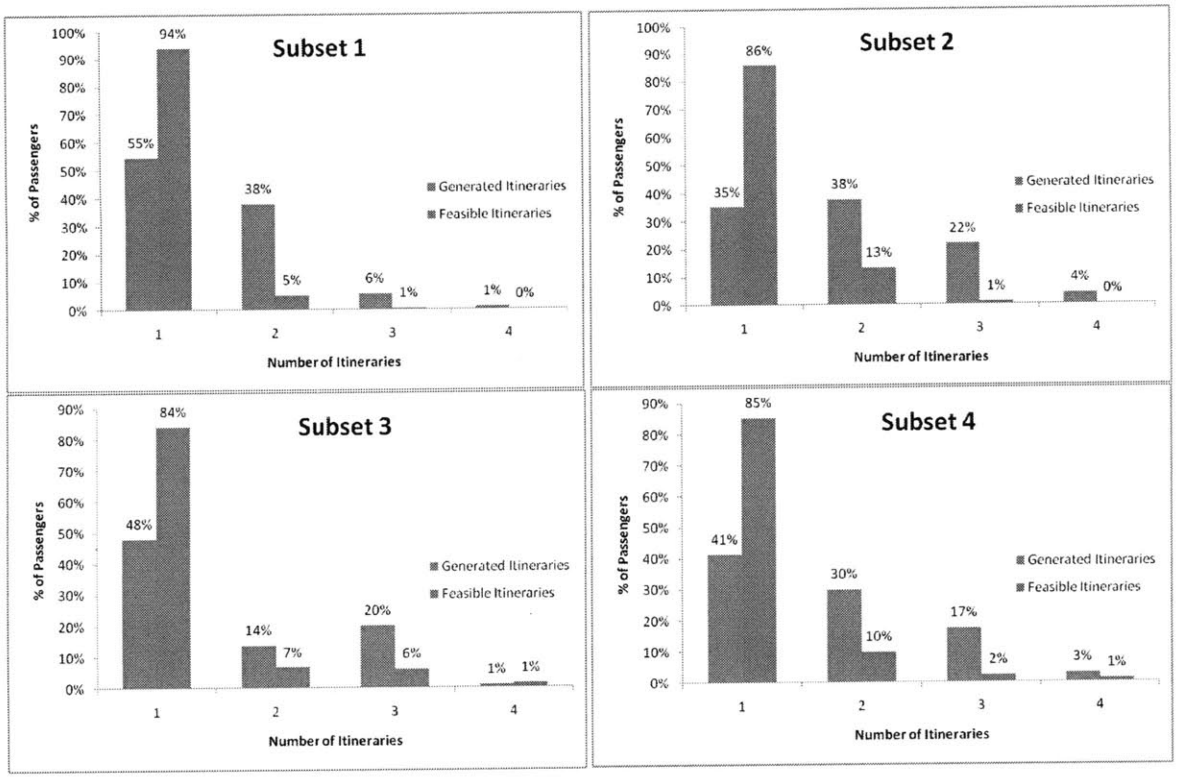
The remaining model results are train- and station-based and cannot be broken down by the subsets defined above. These results are produced for trains and stations on lines with complete NetMIS data, and are shown for the Victoria line as an example. When feasible, the results will be compared to results from LU models to understand the strengths, weaknesses, and remaining questions about this model.

4.2 Number of Possible Itineraries per Passenger

As described in section 3.6, a set of itineraries is generated for each passenger. In section 3.7, two processes are followed to reduce the set of feasible itineraries for each passenger, centered on walk time and route choice:

1. **Walk time:** a passenger will have similar access, egress and interchange times relative to the population of LU riders as a whole.
2. **Route choice:** the route with smallest set of feasible itineraries is likely to be the one chosen by the passenger.

While in some cases the set is reduced to a single itinerary after applying these processes, in many cases there may still be multiple itineraries remaining. Because the model requires that one itinerary is selected, if there is still more than one itinerary, the model applies one last process: the itinerary with the most probable egress time is selected by the passenger. Figure 4-3 shows the number of itineraries per passenger with the original generated set of feasible itineraries, and after the first two processes are applied for each subset of passengers.



4-3 Number of Itineraries for Each Subset Before and After Reduction Processes

Control group: Subset 1, as expected, has the highest percentage of passengers with a single itinerary after the walk time process is applied. Specifically, 94% of passengers have a single itinerary when there is no interference from route choice, interchange and congestion.

Congestion: In subset 2, only 86% of passengers who face congestion have a single itinerary after the walk time assumption is applied. This is to be expected, as passengers who face congestion tend to have longer travel times and therefore more possible itineraries. The model does not perform as well in selecting an itinerary under congestion.

Interchange: In subset 3, 48% of passengers who interchange and experience no congestion between the Victoria and Jubilee lines have a single itinerary after the itinerary generation process. A higher percent of passengers in subset 1 have a single itinerary after the itinerary generation process, but a smaller percent of passengers in subset 2 have a single itinerary after the first step. This is also to be expected because the number of possible permutations of path segments increases with the number of interchanges. 84% of interchanging passengers have one feasible itinerary after the walk time assumption is applied, which is less than the percent for subset 1. These numbers indicate that the model does not perform as well in selecting an itinerary with congestion.

Route Choice: In subset 4, 41% of passengers had a single itinerary before the route choice process was applied. These passengers had multiple possible routes, but the itinerary generation process revealed that only one route was feasible. Itineraries were not generated for the remaining routes because they did not fit within the passengers' travel time for some reason. The remaining 59% of passengers may also have only one feasible route (with multiple itineraries), but they are not discernable from passengers with multiple routes with single itineraries. This means the route choice process is not necessary for at least 41% of passengers with route choice. This is not expected because many OD pairs on the LU network have similar routes that are not too different in terms of travel time, while this result implies that 41% of the OD pairs covered by this subset have routes that are different in travel time. This percentage, while initially promising, may indicate that the model is incorrectly generating too few itineraries for routes. This may be because of the clock misalignment or the timestamp truncation issues. The veracity of this result can be tested by comparing the route choice probabilities from LU's route choice model, with the choice probabilities from this model in future research.

After the route choice and walk time processes were applied, 85% of passengers that face route choice have one feasible itinerary for the chosen route. This percentage is generally in line with that of passengers from other subsets.

4.3 Percent of Passengers Left Behind

As described in section 3.8, a passenger is defined as being "left behind" if he was forecast to be on the platform and did not board the first train that could have taken him to his destination. This section discusses the rate of left behind passengers by subset. It will help identify baseline expectations for left behind passengers, and which factors might distort its estimation. Figure 4-4 shows the percent of passengers left behind for each subset.

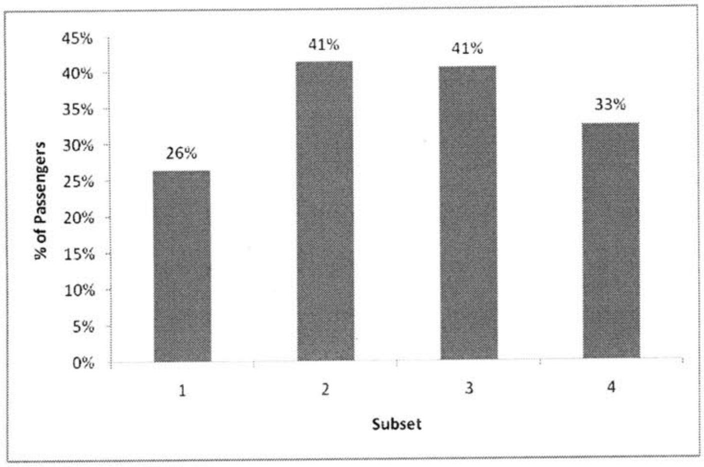


Figure 4-4 Percent of Passengers Left Behind by Subset

Control group: The expectation for subset 1 is that no passengers are left behind because they face no congestion when boarding. Surprisingly, the model classifies 26% of subset 1 passengers as being left behind. This percentage includes the passengers who chose not to board the train for reasons other than congestion, and passengers erroneously classified as being left behind. Few passengers would be expected not to board a train travelling on their desired route, certainly no more than 5% of all passengers. The remaining passengers are erroneously classified as being left behind due to problems in the train itinerary selection process. It is likely that these passengers' access times were estimated incorrectly and they were not on the platform in time to be left behind by the trains in question. This percentage can be viewed as the base error for single leg journeys and is to be expected to be a minimum for any other subset involving single leg journeys.

Congestion: The expectation for subset 2 is that some passengers are indeed left behind because they face congestion. Forty one percent of Subset 2 passengers are left behind. If 26% of these passengers are associated with the base model error, then the remaining 15% of passengers may indeed be left behind because of congestion.

Interchange: The expectation for subset 3 is that no passengers are left behind because they face no congestion, and interchange should have no bearing on being left behind. However, 41% of subset 3 passengers are inferred to have been left behind. Therefore, this percentage includes the passengers who chose to be left behind for reasons other than congestion, passengers erroneously estimated to be left behind when interchange is involved, or a combination of both. The probability for error in left behind estimation increases every time a passenger waits for a train. Because these passengers must interchange, this subset shows a higher percentage of passengers left behind than that of the control group. This percentage can be viewed as the base error for multi-leg journeys.

Route choice: The expectation for subset 4 is that no passengers in addition to the passengers associated with the base error are left behind because they face no congestion, and should choose the route where they are less likely to be left behind. The base error is somewhere between 26% and 41% because this subset includes interchanging and non interchanging passengers. Thirty three percent of subset 4 passengers are left behind. Indeed, the percent of passengers left behind is between the two base error figures. This implies that route choice does not increase the degree of error in estimating left behind passengers. However, in light of the number of feasible itineraries for passengers with route choice, if the wrong itineraries are being generated for passengers, then the wrong itineraries are being selected for these passengers, and they are incorrectly deemed left behind by this model.

4.4 Load on Trains

This section presents the load on Victoria line trains after loading passengers to trains from their selected train itinerary from this model. Figures 4-5 and 4-6 show the inferred load on each Victoria line train from 7:00 to 10:00 AM on May 19, 2009, in the northbound and southbound directions. The trains are sorted by their terminal departure time. Some northbound trains terminate at Seven Sisters, while the rest terminate at Walthamstow Central. Though the loads are uneven from run to run, loads gradually increase over time and then decrease. Figure 4-5 shows northbound peaking from 8:00 to 8:30 AM as expected. Figure 4-6 shows southbound trains on the Victoria line with trains terminating at Brixton. The loads on these trains appear to be more even and peak during the same time period. In general, these loads are what one might expect: low loads during the ramp-up and ramp-down periods (the start and end of the AM Peak), and significant peaking between 8:00 and 9:00 AM. It shows that there is no apparent bias towards some trains over others. In future research, it would be useful to estimate loads for the hour before and after the time period of interest (i.e. from 6:00 to 11:00 AM) so that the loads during the time period of interest would be complete.

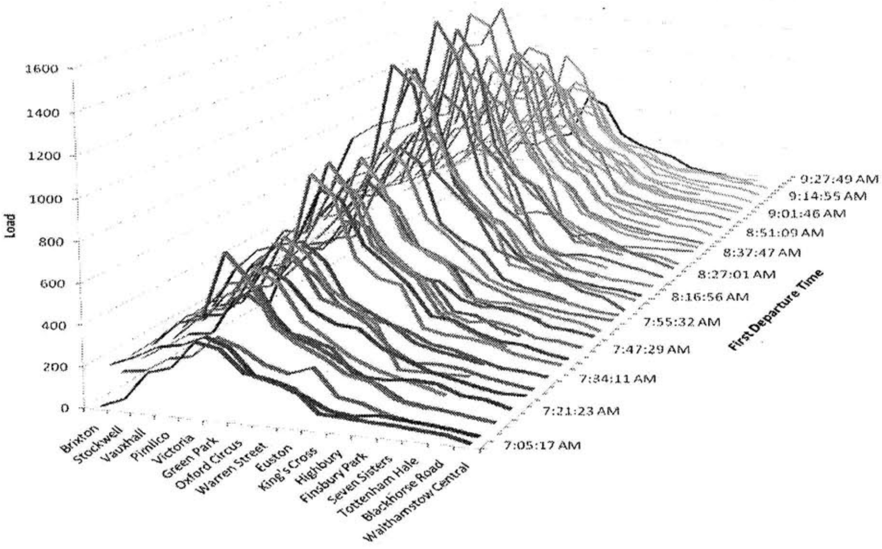


Figure 4-5 Load on northbound Victoria line trains from this model

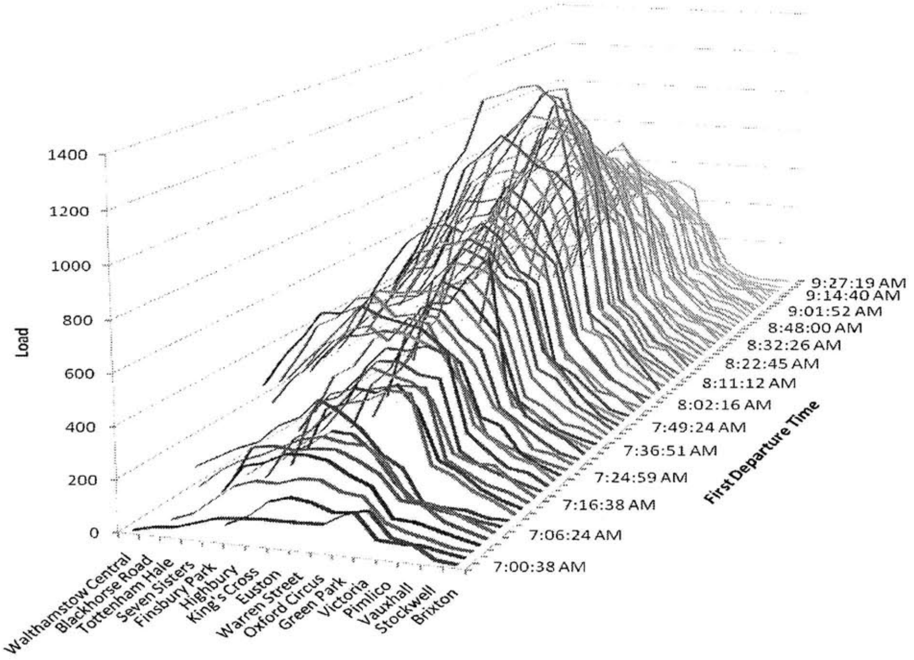


Figure 4-6 Load on southbound Victoria line trains from this model

Load at Individual Stations

The following figure shows the estimated load on each train as it passes through Seven Sisters and Pimlico. Examining the load at individual stations should be useful to study the model's effectiveness in distributing the load across train runs. Seven Sisters was chosen because it is expected to show a high degree of fluctuation in load, while Pimlico is chosen because it is expected to show much less fluctuation. If the estimated load fluctuated at both stations differently than expected, then it would show that the model was not correctly assigning passengers to trains. Figure 4-7 shows two cross-sections of this model's load output from figure 4-6: at Pimlico and Seven Sisters stations, both southbound. Pimlico shows fairly consistent loads as each train departs, while Seven Sisters shows a high degree of fluctuation. This is because some trains start their trip at Seven Sisters, while others served stations north of Seven Sisters prior to arriving at Seven Sisters. The trains that start at Seven Sisters tend to have low loads upon departure, while the trains that started at stations north tend to have much higher loads.

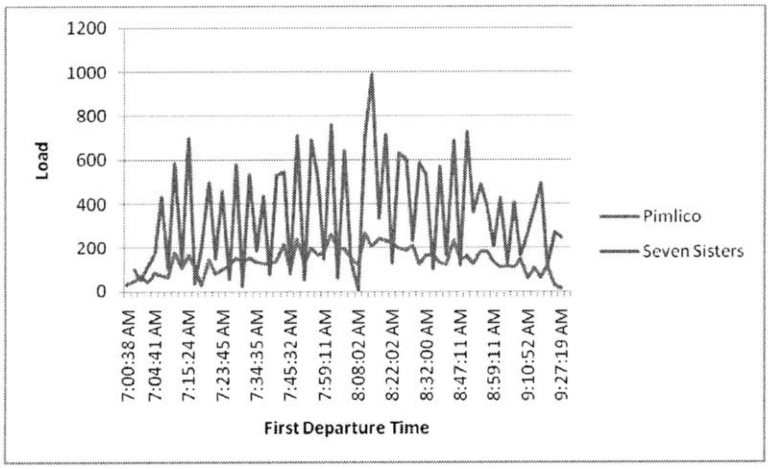


Figure 4-7 Southbound Load at Pimlico and Seven Sisters from this model

Average Load by Time Interval

The following figures show the average load on the Victoria line from this model. Trains are grouped by their departure time in half hour time intervals and their load at each station is averaged. Figures 4-8 and 4-9 are the northbound and southbound loads estimated on the Victoria line. These figures should provide a clearer picture of how load changes by time interval. The peak time interval from this model is 8:00 to 8:30. Another interesting observation is that this model's average loads at certain stations between 8 and 9 AM surpass the crush standing capacity mark, a value specified in the LU Rolling Stock Parameters (Baker, 2009). This shows that when there is no capacity constraint instituted, the model will load passengers onto trains beyond their capacity. This has two possible implications. First, it implies that passengers are incorrectly being assigned to trains and that the itinerary selection process has much room for improvement. The contrary implication is that passengers may be loading themselves onto trains that are full beyond this crush capacity. This is less likely to be true because the crush capacity is a hard physical limit to the number of passengers that can load a train. Future work for this model may involve limiting the load on individual trains at the crush capacity.

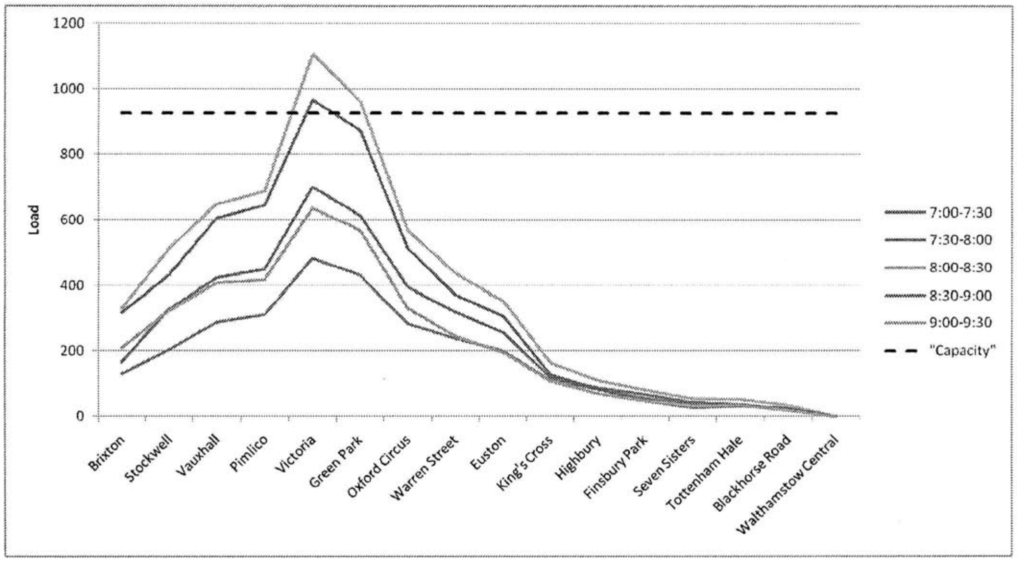


Figure 4-8 Average northbound load on Victoria line by time interval, from this model

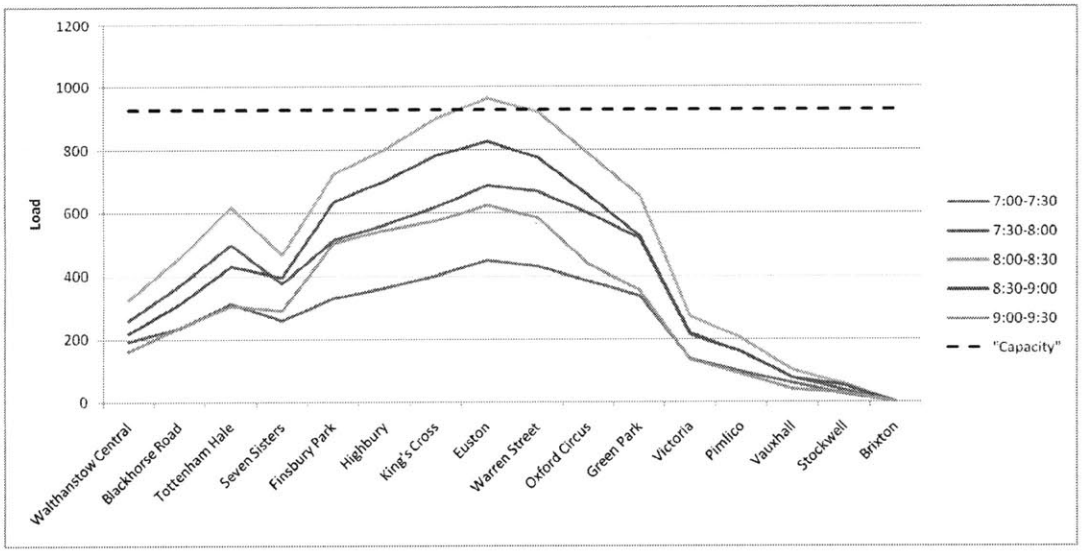


Figure 4-9 Average southbound load on Victoria line by time interval, from this model

4.5 Relationship between Left Behind Passengers and Load

This section explores the relationship between "left behind" passengers (as defined in section 3.8) and load on trains. If this model can reveal the relationship between load and percent of passengers left behind, it will be extremely useful in determining when passengers are left behind. Figure 4-10 shows the relationship between left behind passengers and load on the Victoria line, based on output from this model. The graph excludes any data from trains that did not have complete runs, and started their runs either before 8 AM or after 9 AM. This time restriction is to avoid loads at the beginning and end of the three hour time period which may be lower than reality due to "boundary" issues. For example, loads on trains near the start of the time period may be lower than reality because a number of passengers would have entered the system before the start of the period. The model cannot assign these passengers to trains. Similarly, loads on trains near the end of the time period may be lower than reality because a number of passengers may have exited the system after the end of the time period.

Number of Left Behind Passengers vs. Load

Figure 4-10 shows that there at best is a weak positive correlation between left behind passengers and load. There are just as many instances where trains with high loads have few left behinds, as instances where trains with low loads have many left behinds. This figure also shows that there are many instances where the trains on the Victoria line have loads higher than the crush standing capacity, and the capacity seems to have no impact on whether passengers are left behind or not.

Figure 4-11 on the other hand, shows the relationship between number of left behind passengers and load from a TSM simulation of a typical "no delay" day. No delay is a scenario in which no major incidents occur on the line, though variation still exists within the simulation. For example, if the Victoria line runs without any delays one morning (i.e. no passenger alarms, no signal failures, etc.), it is likely that there is still some delay on the line (i.e. one extended dwell time may interrupt the headways and cause impedance for the trains behind) (Caffull, 2010).This scenario takes into account that trains never run exactly the same each time by randomly instituting minor delays. The no delay scenario was chosen to match with the delay scenario of the morning of May 19, 2010, where there were few incidents.

Figure 4-11 shows that that practically no passengers from TSM are left behind until the load nears the crush capacity and crush capacity has a strong impact on whether or not passengers are left behind. This stark difference stems from the different methods for determining left behinds between the two models. In TSM, passengers are deemed left behind if the load on the train is at (or near) capacity. Otherwise, no passengers are left behind. In this model, a passenger is deemed left behind if the walk time assumptions indicate that the passenger was at the platform before the departure time of the trains that the passenger did not board. The selection of itinerary, much less the decision of considering a passenger left behind, does not consider the load on the set of feasible trains.

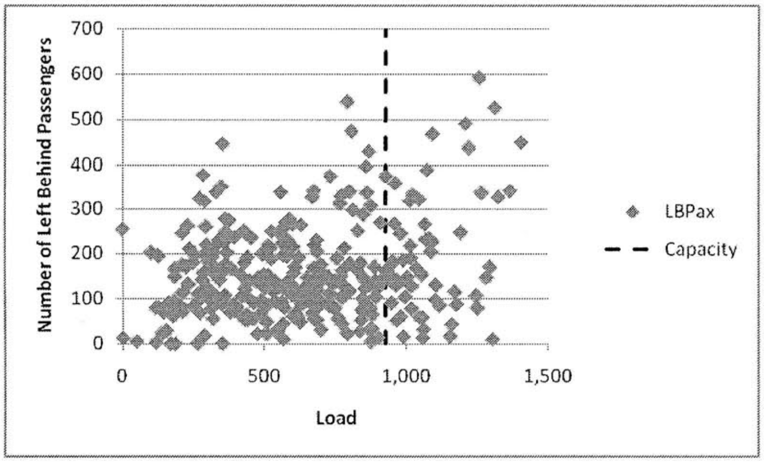


Figure 4-10 Number of Left Behind Passengers vs. Load on Victoria Line, from this model

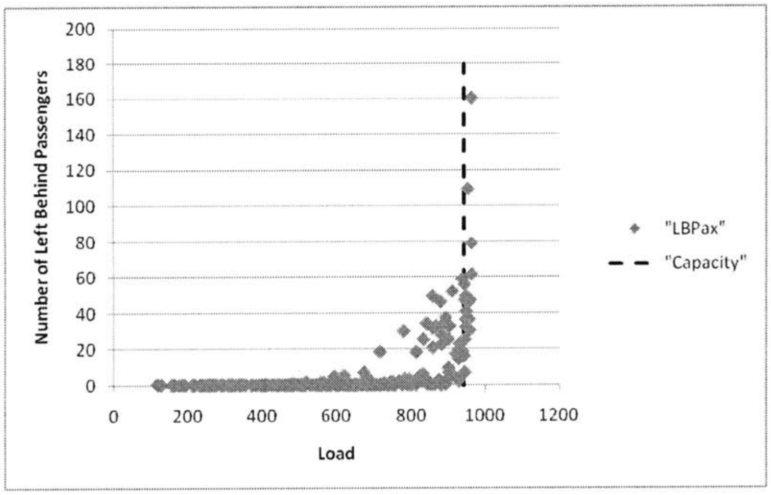


Figure 4-11 Number of Left Behind Passengers vs. Load on Victoria Line, from TSM

Percent of Left Behind Passengers vs. Load

A more useful way of looking at the relationship between load and left behind passengers is a graph of load versus the ratio of left behind passengers to the total number of passengers attempting to board a train. Figures 4-12 and 4-13 describe this relationship. In figure 4-12, the relationship between load and percent of passengers left behind, output from this model, appears to be completely random. Crush capacity has no bearing on this relationship. In some instances, the percent of passengers left behind is as high as 70%. Incidentally, the results from section 4.3 indicate that 26% of passengers on single leg journeys that are marked as left behind were done so in error. This error is certainly apparent in this figure. In figure 4-13, the relationship between load and percent of passengers left behind, output from TSM, is similar to figure 4-11. The maximum percent of passengers left behind is no more than 45%.

Figures 4-12 and 4-13 represent two extremes in counting left behinds: this model does not take load into consideration, while TSM allows load to be the sole factor in determining left behinds. Neither approach reflects the expected relationship between left behind passengers and load. While most passengers should be left behind when load is at capacity and few passengers should be left behind when load is low, some passengers should be left behind as the load approaches capacity. In future research, if the itinerary selection process is improved, it should reflect this relationship.

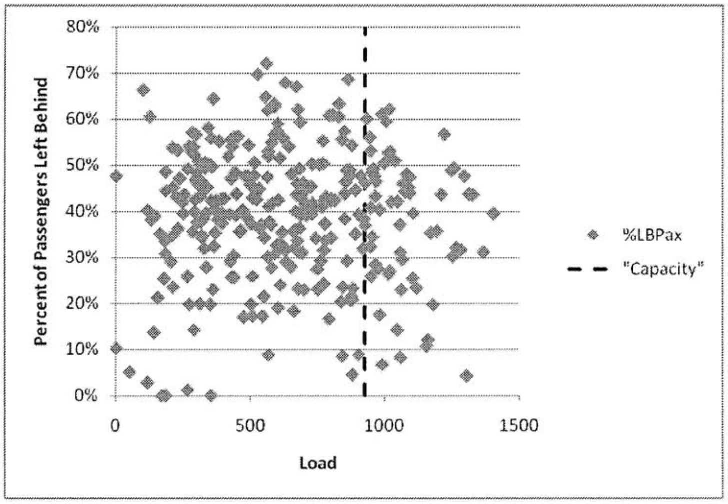


Figure 4-12 Percent of Passengers Left Behind vs. Load on Victoria Line, from this model

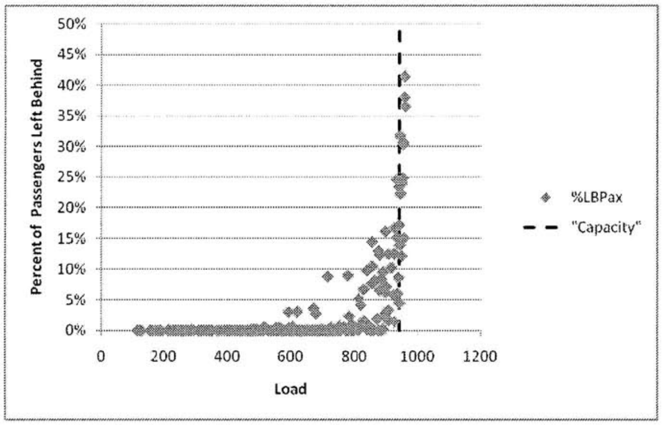


Figure 4-13 Percent of Passengers Left Behind vs. Load on Victoria Line, from TSM

4.6 Exploration of Assumptions

In this section, three of the assumptions discussed in chapter 3 will be modified to see if they improve the results for the model:

- Oyster time stamp assumption.
- Walk time assumption.
- Route choice assumption

These assumptions are tested by generating the number of feasible itineraries for each subset under different assumptions, and comparing those figures to those generated under the assumptions described in chapter 3. Figure 4-3 shows these figures: Subset 1 starts with 55% of passengers with one possible itinerary after the itinerary generation process, and increases to 94% after itinerary reducing assumptions are applied. Subset 2 goes from 35% to 86%. Subset 3 goes from 48% to 84%. Subset 4 goes from 41% to 85%.

4.6.1 Oyster Timestamp Assumption

The truncation of Oyster timestamps is discussed in section 3.2.1: Oyster records the minute for each transaction, but not the seconds. The model assumes the narrowest possible window between Oyster tap-in and tap-out: the entry time is assumed to be at the end of the minute, and the exit time is left as is. In this section, the model generates results based on the widest window assumption: the entry time is left as is, and the exit time is assumed to be at the end of the minute. This assumption goes into effect prior to the itinerary generation process. Figure 4-14 presents the resulting number of itineraries generated for passengers in each subset. In all subsets, the results have worsened significantly: all passengers have more possible itineraries. As a result, after the itinerary reduction process is applied, fewer itineraries are eliminated. Under the modified timestamp assumption, subset 1 has only 81% of passengers with one possible itinerary, subset 2 67%, subset 3 47% and subset 4 64%.

The fact that the results have not improved indicates that the original timestamp assumption is the better assumption. It also suggests that this model is very sensitive to changes in the travel time window allowed for passengers. This highlights the seriousness of the Oyster clock misalignment issue, discussed in section 3.2.1. Though the clock misalignment is believed to be on the order of a few minutes, it could have a significant impact on the number of itineraries generated for each passenger. This suggests that the ability of this model to estimate train loads may improve after the clock misalignment issue is resolved.

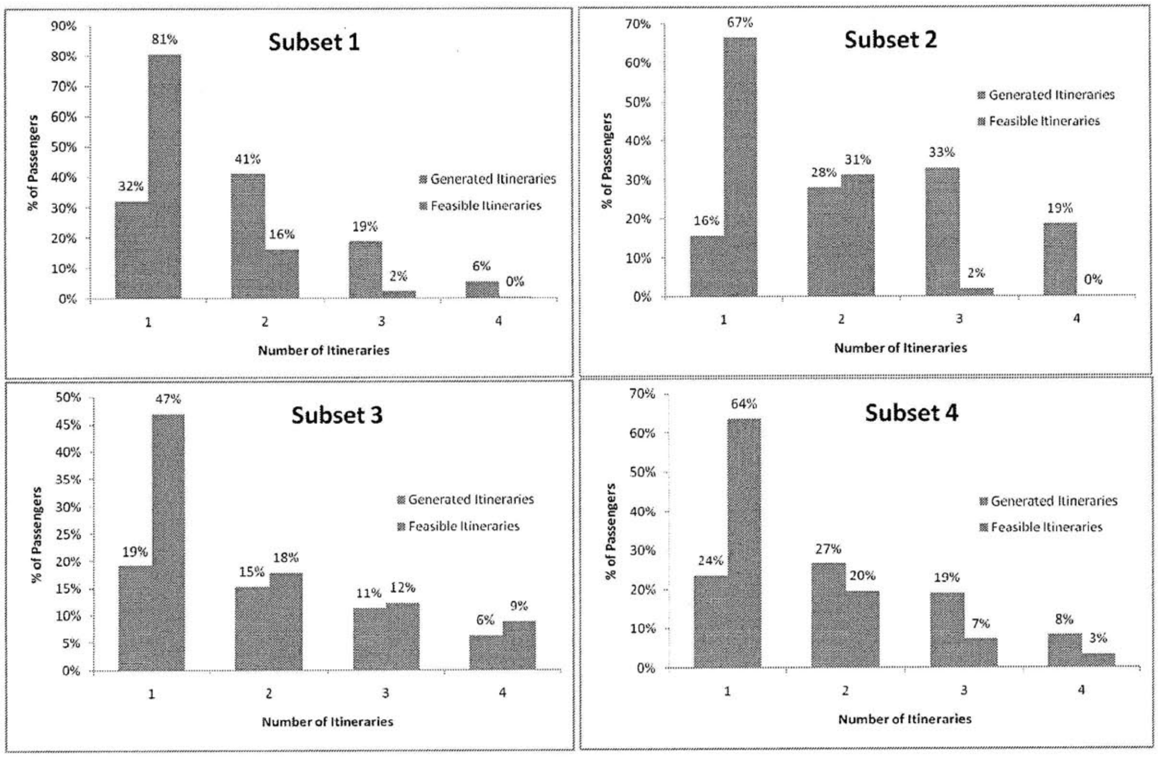


Figure 4-14 Number of Feasible Itineraries for Each Subset under new timestamp assumption

4.6.2 Walk Time Assumption

The walk time assumption discussed in section 3.7 is used to reduce the set of feasible itineraries. Passengers are assumed to have similar access, egress and interchange times. To account for this, itineraries that do not allow for similar walk times are eliminated. In this section, a different approach is taken to the relationship between walk times and the set of feasible itineraries. If all passengers are assumed to have the average walk times based on LU's standard values, then itineraries that do not allow for average walk times are eliminated. Figures 4-15 and 4-16 show the resulting number of feasible itineraries for each passenger in subsets 1 and 2. The results are significantly worse than the original results. In subset 1, 41% of passengers have no feasible itineraries because of this constraint. This means that the constraint was too restrictive on those passengers. In subset 2, the constraint was too restrictive on fewer passengers (32%). Regardless, this modified assumption is not as good as the original because it eliminates too many feasible itineraries. This suggests that the relationship between walk times and the set of feasible should be treated more delicately. For this reason, the walk time assumption used in this model is better than the one presented in this section. The other subsets are not included because they do not provide any new information.

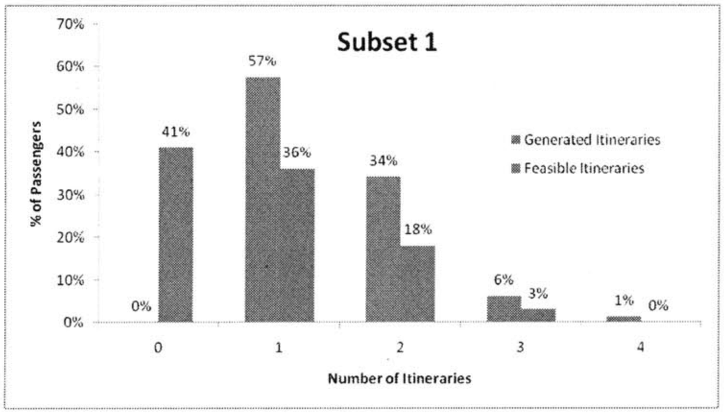


Figure 4-15 Number of feasible itineraries for Subset 1 before and after modified walk time assumption

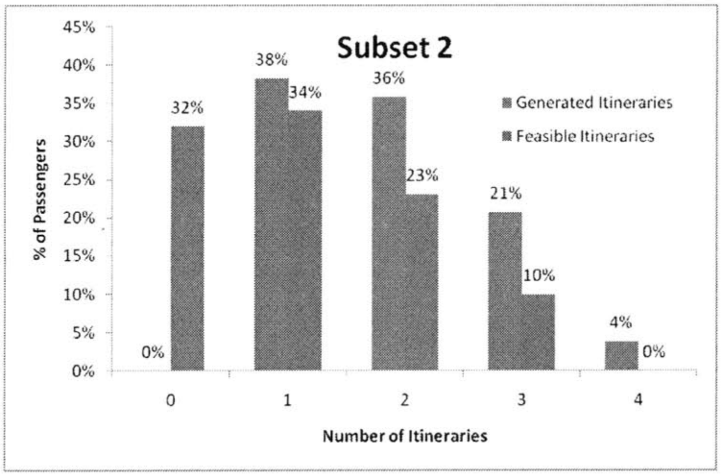


Figure 4-16 Number of feasible itineraries for Subset 2 before and after modified walk time assumption

4.6.3 Route Choice Assumption

The route choice assumption discussed in section 3.7 is used to reduce the set of feasible itineraries. For passengers that have a route choice, a route is chosen prior to the application of walk times to eliminate feasible itineraries. The route with the fewest feasible itineraries is chosen. The walk time assumption is then applied to these itineraries. In the alternative approach, the route will be chosen prior to the generation of itineraries using a different criterion: generalized cost. Routes are sorted by generalized cost, which is a function of utility described in section 3.2.7. The route with the least generalized cost is considered first. If no feasible itineraries can be generated for this route, then the next route on the list is considered. If the route under consideration has feasible itineraries, no other routes are considered. Figure 4-17 shows the resulting number of feasible itineraries for passengers in subset 4, the only subset that involves route choice. The results have significantly worsened compared to the original subset 4 results. Only 29% of passengers have one feasible itinerary after the itinerary generation process. The results are worse because this method for selecting a route does not take each passenger's travel time into consideration. This method gives preference to routes that are shorter in travel time and simpler in terms of number of interchanges. However, if a passenger has a longer travel time, many feasible itineraries will be generated for the short route. For this reason, it is better to consider route choice after the itineraries are generated but before itinerary reduction assumptions are applied. This method accounts for passengers' travel time. In light of the possible error with the current route choice process, another approach may be to choose routes probabilistically (using equation 2-3). This should be explored in future research.

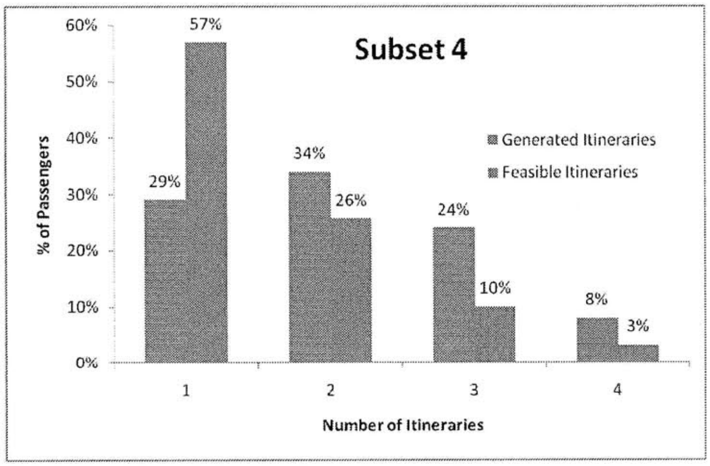


Figure 4-17 Number of feasible itineraries for Subset 4 with modified route choice assumption

5 Research Summary & Conclusions

This chapter begins with a summary of this research in section 5.1, including a description of the objective and the framework built to achieve the objective, and an overall assessment of this research. Section 5.2 evaluates the model developed in this thesis by describing the challenges faced, and assessing its methods in addressing the challenges. Section 5.3 summarizes the research findings and assesses the model's effectiveness in achieving the research objective. Lastly, section 5.4 identifies future directions for continuation of this research.

5.1 Summary

This thesis attempted to assess the possibility of identifying which train a passenger took to get from his origin to destination while travelling in a high frequency urban rail transportation system. It reviewed previous attempts at solving this problem, and developed a model that draws from the lessons from these previous works and attempts to solve this problem in the context of the London Underground.

The model estimated the passenger demand and the service supply in the London Underground network through the use of smartcard and track signaling data, and a route choice model. These data sets were used to identify passengers' train level itineraries through a temporal and spatial matching process. When passengers had multiple feasible train itineraries, a series of assumptions on route choice and walk times were applied. Once the train itineraries were identified, the loads on trains, number of passengers left behind by trains, and access, egress and interchange distributions were presented. Finally, these figures were compared to estimates from London Underground. The model was designed in the context of the London Underground, but can be applied to other urban public transportation systems.

Assessment of the results and consideration of the challenges in the creating the model does not conclusively indicate that identifying the exact train a passenger selects to get from his origin to destination is possible, at least in the case of the London Underground. However, the results do indicate that the model has significant potential, and can be improved in future research. These initial results can serve as indicators on how to improve the model.

5.2 Model Evaluation

In the process of developing the model, many challenges arose. This section will summarize these challenges and evaluate the model's effectiveness in overcoming these challenges. The challenges are as follows:

- **Oyster Clock Misalignment:** The clocks which supply the Oyster transaction times at fare gates in stations across the LU network may not be synchronized. Therefore, passengers' entry and exit times may be incorrect, which may affect the accuracy of the model in assigning passengers to train itineraries. The model attempts to overcome this challenge by increasing a passenger's travel time in one-minute increments if the clock misalignment has caused the passenger's travel time to be too short. However, if the clock misalignment has caused the travel time to be too long, the model makes no adjustment, and the additional time a passenger appears to have spent in the system is considered to be walk time or platform wait time. There is room for improvement in adjusting for too long travel times.
- **Missing Data from NetMIS:** NetMIS is an incomplete data set. Out of the 11 lines on the London Underground, only 3 have completely reliable data. The algorithms the model employs to adjust for the missing data largely depend on expected run time, with a two minute margin of error. This research has not tested the degree to which NetMIS run times are in line with the expected run time. Also, this two minute margin of error is an arbitrary value. There is room for improvement in how this model handles run times.
- **Oyster Timestamp Truncation:** The Oyster dataset only reports the minute of entry and exit transactions. This lack of precision creates ambiguity in a passenger's travel time. Faced with two possible extremes, the model assumes the shortest possible travel time, which means that while no infeasible itineraries will be generated for passengers, some feasible itineraries may be excluded. In the tradeoff between feasibility and inclusiveness, the model elects to generate only feasible itineraries. This model's general approach to this tradeoff is to favor feasibility to increase the accuracy of the model, as long as measures are taken to account for too much exclusiveness. In the case of timestamp truncation, if assuming the shortest possible travel time produces no feasible itineraries, the travel time is increased in one minute increments until feasible itineraries are generated.
- **Uncertainty and inability to validate results:** When the set of feasible itineraries generated for each passenger contains more than one itinerary, there is no certain way to determine which

itinerary is actually selected. In order to overcome this challenge, the model assumes that passengers have similar access, egress and interchange times, and eliminates itineraries that do not allow for that quality. While this assumption is effective in reducing the size of the set of feasible itineraries, it may also eliminate the actual itinerary of passengers if they have dissimilar access, egress and interchange times. Again, the model favors feasibility over inclusiveness, but attempts to account for too much exclusiveness.

5.3 Research Findings

This section summarizes the results of the application of model is applied to the London Underground. The model reports the number of feasible itineraries, the percent of passengers left behind, passenger load on each train, the relationship between loads and left behind passengers, and average egress times.

- **Number of Feasible Itineraries:** The model reports the number of feasible itineraries after the itinerary generation process (the first stage) and after the walk and route choice assumptions are applied (the second stage). Ideally, most passengers would have one feasible itinerary after the first stage. The model is determined to be accurate in selecting itineraries based on the percent of passengers that have one feasible itinerary after the first and second stages. The following results were obtained for different subsets of passenger trips, as defined in section 4.1:
    - **Subset 1:** When there is no interference from congestion, route choice and interchange, the model performs well. It finds a single itinerary for the majority of passengers after the first stage, and nearly all passengers after the second stage.
    - **Subset 2:** The model does not perform as well when congestion exists.
    - **Subset 3:** The model performs worse in selecting the itinerary when interchange is involved. This is because the number of possible permutations of path segments increases with the number of interchanges.
    - **Subset 4:** When route choice is involved, the model is able to determine with certainty the route of 41% of passengers after the first stage—before the route choice assumption is applied. This result is questionable and may indicate that the model does not handle route choice well.

- **Percent of Passengers Left Behind:** The model reports the percent of passengers estimated to be left behind by trains for each subset. The control subset, where passengers do not face any congestion, has 26% of passengers left behind. This means that the base error in determining left behind passengers for single leg journeys is quite large. After subtracting the base error, the model determined that 15% of passengers on single leg journeys are left behind due to congestion. The base error for multi-leg journeys is even larger.
- **Passenger Load on Trains:** The model computes the passenger load on each NetMIS train on May 19, 2009. The load appeared to be as expected, peaking during the right time periods. However, the load estimated from this model on occasion also exceeded the crush capacity constraint, which is an indication that passengers may be assigned to the incorrect train itineraries.
- **Relationship between Left Behind Passengers and Loads:** It has already been noted that left behind passengers are over estimated for all groups of passengers. There is a positive correlation between left behind passengers and load, but relationship was not as strong as expected. The crush capacity seems to have no impact on whether passengers are left behind or not. This is another indication that passengers are incorrectly determined to be left behind.
- **Average Egress Times:** This model compares the average egress time at each station to LU's standard egress time in section 3.7.1. Nearly all stations manually surveyed by LU appear to be in line with this model's egress time. This indicates that this model accurately estimates egress times, and stations not manually surveyed by LU may have underestimated standard egress values.

In summary, the model performs well in generating itineraries, especially with simple journeys. However, there are indications that the correct itineraries are not selected for some passengers. Load estimations appear to be similar to load estimated in LU models, with some small differences. Left behind passengers are overestimated, which in turn would cause the relationship between load and left behind passengers to be weak.

5.4 Future Research

This section highlights the most important additional work that can be developed by future researchers. There are several threads of research stemming from this work that future researchers could find useful to explore. These points are mainly extensions to this research and ideas that will help improve the accuracy of the model, and help determine if it is possible to identify the exact train a passenger selects.

- **Iterative process:** A way to improve this model could be to make it an iterative process. Load estimates from previous iterations could be used to inform the selection of itinerary for passengers, and the determination of a left behind passenger. Likewise, access, egress and interchange time distributions for all stations from previous iterations could also be used to inform the selection of itinerary.
- **Probabilistic approach to walk time assumption:** A more complex approach to passenger walk times could be developed. Although the majority of passengers might have similar access, egress and interchange times, this may not always be the case. An approach that would select the itinerary that maximizes the probability that a passenger has similar walk times (instead of eliminating itineraries that do not allow for similar walk times) may improve the accuracy of the model.
- **Access, Egress and Interchange time distributions by station and line:** Walk time distributions may be improved if they are measured and reported to a greater level of detail: by station and line. There appears to be sufficient data for this kind of analysis to be successful. In the interest of time, this idea was not implemented.
- **Computing the choice probabilities of routes**: For OD pairs that have multiple possible routes, the results from this model can be used to compute the probability of passengers choosing each route in the path choice set. These probabilities could be used to improve other models in LU as well as aid in the itinerary selection process in this model.
- **Other time periods:** This model could be run during other time periods, particularly during time periods where there is less passenger demand and lower frequency of trains. This scenario may allow researchers to study how the generation of itinerary process differs when passenger demand and service supply are different.
- **Exploring the relationship between left behinds and load among subsets of passengers:** An interesting study may be to separate passengers by trip length and explore the differences in the percent of passengers left behind vs. load for each group.

Bibliography

Baker, C. (2010), Unpublished interviews and electronic communications with Charles Baker, Senior Planner, London Underground, Strategy and Service Development.

Barry, J. J., Newhouser, R., Rahbee, A., and Sayeda, S. (2001), Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record,* 1817:183-187.

Ben-Akiva, M. and Lerman, S. (1985), Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press.

Buneman, K. (1984), `Automated and passenger-based transit performance measures', *Transportation Research Record 992*, 23-28.

Cafful, K. (2010), Unpublished interviews and electronic communications with Kirsty Cafful, Planner, London Underground, Strategy and Service Development.

Chan, J. (2007), Rail OD matrix estimation and journey time reliability metrics using automated fare data, Master's thesis, Massachusetts Institute of Technology.

Frumin, M. (2010), Automatic Data for Applied Railway Management: Passenger Demand, Service Quality Measurement, and Tactical Planning on the London Overground Network, Master's thesis, Massachusetts Institute of Technology.

Gami, R. (2010), Unpublished interviews and electronic communications with Rajesh Gami, Planner, London Underground, Strategy and Service Development.

Gordillo, F. (2006), The Value of Automated Fare Collection Data for Transit Planning: An Example of Rail Transit OD Matrix Estimation, Master's thesis, Massachusetts Institute of Technology.

Guo, Z. (2008), Transfers and Path Choice in Urban Public Transport Systems, PhD thesis, Massachusetts Institute of Technology.

Kusakabe,T., Takamasa, I., Asakura, Y. (2009), Estimation Method for Railway Passengers' Train Choice Behavior with Smart Card Transaction Data, Kobe University Graduate School of Engineering.

Nuzzolo, A., Russo, F. and Crisalli, U. (2001), `A doubly dynamic schedule-based assignment model for transit networks', *Transportation Science* 35(3), 268-285.

Nuzzolo, A. and Crisalli, U. (2009),'The Schedule-based modeling of transportation systems: recent developments', *Schedule-Based Modeling of Transportation Networks: Theory and Applications*, Operations Research/Computer Science Interface Series, Spring Science+Business Media, chapter 1, pp. 1-26.

Prashker, J. and Bekhor, S. (2004), `Route choice models used in the stochastic user equilibrium problem: A review', *Transport Reviews* pp. 437-463.

Rahbee, A. (2010), Unpublished interviews and electronic communications with Adam Rahbee, formerly of London Underground Strategy and Service Development.

Roberts, M. (2010), Unpublished interviews and electronic communications with Mark Roberts, Transport for London, Fares and Ticketing.

Spiess, H. and Florian, M. (1989), `Optimal strategies: A new assignment model for transit networks', *Transportation Research B* 23(2), 83-102.

Transport for London (1999), 'Journey Time Metric (JTM): An Overview'.

Transport for London (2009a), Business Case Development Manual.

Transport for London (2009b), `Oyster Factsheet'. http://www.tfl.gov.uk/assets/downloads/corporate/oyster-factsheet-july-2009.pdf (Accessed April 17, 2010).

Uniman, D. L. (2009), Service reliability measurement framework using smart card data: Application to the London Underground, Master's thesis, Massachusetts Institute of Technology.

Weston, G. and Maunder, G. (1994), 'Train Service Model – Technical Guide', Operational Research Note 89/18.

Weston, G. (2009), Unpublished interviews and electronic communications with Gerry Weston, London Underground, Strategy and Service Development.

Weston, G. (2010), Unpublished interviews and electronic communications with Gerry Weston, London Underground, Strategy and Service Development.

Wilson, N.H.M.W, Zhao, J., Rahbee, A.,(2009), 'The potential impact of automated data collection systems on urban public transport planning', *Schedule-Based Modeling of Transportation Networks: Theory and Applications*, Operations Research/Computer Science Interface Series, Spring Science+Business Media, chapter 5, pp. 75-97.